

Machine Learning Analysis of the Cultural and Cross-Cultural Aspects of Beauty in Music

Claire Q

17/05/13

Contents

1	Introduction	1
2	Machine Learning	6
2.1	Introduction	6
2.2	Training	7
2.2.1	Supervised Learning	9
2.2.2	Unsupervised Learning	11
2.2.3	Active Learning	12
2.2.4	Semi-supervised Learning	13
2.2.5	Reinforcement Learning	14
2.3	Learning Algorithms of Interest	15
2.3.1	Bayesian Reasoning	15
2.3.2	Decision Trees	17
2.3.3	Artificial Neural Networks	17
2.3.4	Support Vector Machines	19
2.3.5	Inductive Logic Programming	20
2.3.6	k -Nearest-Neighbour algorithms	21
2.4	Machine Learning Tools	22
2.4.1	Example ARFF file	24
3	Machine Learning, Aesthetics and Musical Aesthetics	27
3.1	Machine Learning and Aesthetics	27
3.2	Musical Aesthetics	30
3.3	Machine Learning and Music	34
3.3.1	Source Material and Representations	40
3.3.2	Audio Feature Extraction Tools	48
3.3.3	Areas of Interest for Classification	49

3.3.4	Geospatial Analysis	66
4	Beauty Experiments	69
4.1	Introduction	69
4.2	Facebook Survey	70
4.3	Results	73
4.4	Learning Algorithm Experiments	75
4.4.1	Experiment Environment	75
4.4.2	Richard Thompson Detector	76
4.4.3	Benchmarking	78
4.4.4	Design	78
4.5	Last.fm	84
4.5.1	Method	84
4.5.2	Result and Discussion	85
4.6	Conclusion	86
5	Geographical Experiments	87
5.1	Introduction	87
5.2	Geographical Ethnomusicology	88
5.3	Method	90
5.3.1	Music Collection	90
5.3.2	Audio Features	93
5.3.3	Geographic Representation	94
5.3.4	Spherical k -Nearest-Neighbour Prediction Method	95
5.3.5	Utilising <i>a priori</i> Background Knowledge	97
5.4	Results	100
5.4.1	k -Nearest-Neighbour Performance	100
5.4.2	kNN with Population Distribution	101
5.4.3	Statistical Significance	104
5.4.4	Performance by Country	105
5.5	Discussion and Future Work	107
5.6	Summary	109
6	Beauty in World Music	113
6.1	Introduction	113
6.2	Design	114
6.2.1	Music Used	114
6.2.2	Listening Conditions	115

6.2.3	Participation	116
6.2.4	Grouping of Listeners and Pairs	118
6.3	Survey Results	118
6.3.1	User Demographics	119
6.4	Learning Beauty	124
6.4.1	Introduction	124
6.4.2	Method	124
6.4.3	Results	125
6.4.4	Discussion	126
6.5	Conclusion	126
7	Beauty and Geography	128
7.1	Introduction	128
7.1.1	Statistical Dependence on Geography	129
7.1.2	Distance from Singapore as a Predictor of Beauty Rating	130
7.1.3	Results	132
7.2	Experiment 2: Location and Audio	134
7.2.1	Method	134
7.2.2	Results	134
7.3	Conclusion	135
8	Conclusion	137
8.1	Summary of Work	137
8.2	Initial Experiments	139
8.3	Geographic Prediction	139
8.4	Singapore Survey	140
8.5	Geographical Influence on Ratings	140
8.6	Future Work	141
8.7	Remarks	142
A	Full Results from Singapore Survey	144
B	User Demographic Data for the Singapore Survey	152

Abstract

Can machine learning algorithms be trained to recognise beauty in music? To what extent is human recognition of beauty in music cultural, or cross-cultural?

Music is prevalent in all human cultures. Music information retrieval is a growing field in which computational techniques have been applied to many musical problems such as genre recognition and measuring musical similarity. Computational ethnomusicology is rarer because the acquisition of non-Western music is difficult. Beauty in music has been little investigated with scientific methods, though there are some examples on which this thesis builds.

The effect of timbral and 12-step chroma audio features, and a wide variety of different machine learning algorithms techniques were tested, with the combination of all the MARSYAS features and Support Vector Machines performing well.

Predicting beauty was first investigated with a small Last.fm set and later with a larger world music survey with Singaporean participants. Beauty was predicted based on a small selection of Last.fm tags with good accuracy. Beauty ratings from the survey, conducted in Singapore, were predictable by machine learning using similar methods.

Predicting the geographical origin of world music from audio features was attempted. Some promising results emerged, and novel methods for predicting points on the surface of the Earth were developed.

An investigation into the link between beauty ratings and location was conducted. The Singaporean beauty ratings were predicted from audio content, geographic content and a combination of both, showing strong correlations between longitude, distance, and timbral features with the beauty ratings, which were statistically very closely linked with distance from Singapore. From this beauty in music is concluded to be culturally related and timbre is shown to be a good pointer to cultural differences.

Acknowledgements

This work was made possible by EPSRC via the Computer Science Department in Aberystwyth, whose funding is highly appreciated.

The creators of MARSYAS and Weka are owed a debt of gratitude.

Thanks go to all in the Dept who have had input, especially my supervisors Professor Ross King, Dr Maria Liakata, and Professor Mark Neal. Others who have been invaluable and supportive include Professor Qiang Shen for unwavering support, Horst Holstein who advised on geographic mathematics, Edel Sherratt for operations research and coffee room banter, Andy Starr for coffee-room cynicism, Richard Shipman for introducing me to LARP which provided valuable escapism, Kevin Williams for sarcasm, unleashing the moon, and free cake, Roger Boyle and Hannah Dee for camaraderie and seemingly boundless hospitality. Thanks to everyone else in the department, especially those with whom I've shared an office or a drink.

Thanks to all my friends near and far who have supported me and endured my stress levels, my Dad, for persistent encouragement, and lastly and most importantly my partner Jimmy Carter, who in addition to being generally lovely somehow found time to dig his way out of his own thesis to read mine.

Dedicated to the memory of my mum, Joy, who would say 'I always knew you could do it'.

Chapter 1

Introduction

Music has been part of the human experience for tens of thousands of years [McDermott and Hauser, 2005]. Musical composition, performance, dancing, singing and listening permeate all cultures and all generations. Beautiful music is appreciated by any listener, not just experts of musicology. With the recent progress in machine learning and in music information retrieval comes an opportunity to investigate beauty in music more scientifically than ever before, to shed light on the nature of beautiful music and to discover how appreciation of music is intertwined with physics, psychology, culture and geography.

Music can be concordant or discordant; this is known from the physics of wave propagation, and the note systems which have emerged across the world reflect this. Very discordant sounds are perceived negatively by any human

and, indeed, by some other primates [Sugimoto et al., 2010]. Expert knowledge can say only so much since expertise tends to be for a particular genre of music, or for Western music theory, much of which does not apply to music from other parts of the world. The scales we use are a compromise for keyboard instruments and the restriction does not hold elsewhere.

Beauty is the subject of many essays and arguments, yet few scientific investigations. Some analysis has been done on visual art which shows, for example, that symmetry plays a part in the beauty of human faces as judged by others. It is possible to computationally separate a Jackson Pollock (very abstract art) from an imitation. Aesthetics in many art forms, including music, has been posited to align with power laws in the use of particular elements, such as the use of notes and rhythms in music or colours in paintings. Fractals have been used to generate art in many forms, never yet surpassing human creations and indeed involving a great deal of human intervention or 'tuning'. Clearly we have a way to go before the first computer-generated top-ten hit.

Music information retrieval is a growing field with focus on automatically extracting information from musical sources for analysis. The musical source comes in many formats including written score as well as audio. A variety of machine learning and statistical analysis techniques are applied. Work in the field of music information retrieval has discovered features for predicting genre, determining key and tempo of music, distinguishing instruments (still

difficult for orchestral music), analysing the similarity of music, transcribing to score from audio and eliciting musical information from written scores. Yet, much of this work has focussed on Western music alone. Investigations have begun into how applicable current techniques are to non-Western sources.

Non-Western music often has a different tuning, sometimes is not strictly tied to a particular tuning, often scales are chosen from a set of possible notes greater than Western music allows (twice in the case of Arabic music), the instruments used are often unfamiliar to the Western listener. Aside from inherent musical differences the production quality of much world music available is not good, and it can be difficult to find large corpuses of non-Western music for analysis. The metadata available about non-Western music is often much sparser than for Western music. These factors probably influence the lack of machine learning analysis on world and ethnic music until fairly recently.

With this in mind the overarching hypothesis is that beauty in music is measurable and is not entirely dependent on cultural and individual factors, but has some root in the commonalities of how the human ear and brain process sound. If the measure is to be objective one way to achieve this is using machine learning. Others have successfully applied machine learning to various music information retrieval problems, and so it is in their footsteps we follow.

This thesis is arranged as follows:

- Chapter 2: A review of the field of machine learning. Supervised, unsupervised and semi-supervised learning are described, including several specific algorithms such as Support Vector Machines (SVMs) and k-Nearest-Neighbour algorithms.
- Chapter 3: A review of Music Information Retrieval (MIR) literature. Common applications and areas of interest in MIR are described, useful features and recent developments are considered. Associated computational aesthetics and non-computational beauty literature is discussed.
- Chapter 4: A survey conducted on Facebook to determine the feasibility of humans agreeing about beauty in music. Participants choose which of two classical extracts is most beautiful. Genre benchmark testing experiments comparing different machine learning algorithms on a known task and dataset, and related investigations. Learning beauty from online user tags in Last.fm.
- Chapter 5: An investigation into the prediction of world music origin using machine learning. kNN regression is applied to world music audio features in order to predict the location of origin as a point on a globe. A land mask is added to improve the prediction, and a population overlay. Predictions are of spherical distance, which is novel for this area.

- Chapter 6: A survey conducted in Singapore on the beauty of the world music of Chapter 5 as rated by participants. Audio features from this music are used to predict the beauty ratings with SVM.
- Chapter 7: An investigation of the degree to which the ratings from Singapore were affected by the location of origin of the music. The ability to predict the ratings from geographic information alone is trialled.
- Chapter 8: A summing-up of all the experiments, results, and any insights that can be drawn. An evaluation of the work achieved and a suggestion of some next steps for further work, including expanding beauty experiments and gathering more world music and different participants.

Chapter 2

Machine Learning

2.1 Introduction

Machine learning is a broad field, covering many kinds of learning problems and learning methods, with several aims including classification and prediction, problem solving, data mining, natural language processing, speech recognition, and visual interpretation [Mitchell, 1997]. The applications of machine learning are many and varied, from creating realistic opponents or allies in video games to making robots for space missions that can automatically plan and execute their actions with little intervention from their creators, to music recommender systems for online listeners. Machine learning can be considered a subset of artificial intelligence research, though the goals of both AI and ML have changed from making a generally intelligent

machine to creating useful algorithms with just the right kind of intelligence for the intended application. Artificial intelligence also encompasses planning algorithms, machine perception and robotics.

This part of the literature review comprises a general look at the machine learning field, later focusing on those techniques which are most appropriate and which became most useful in the experiments conducted.

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T as measured by P improves with experience E .” [Mitchell, 1997]

2.2 Training

The strength of a machine learning agent is strongly dependent on the quality of the data on which it is trained. Noisy examples inevitably lower performance [Nettleton et al., 2010, Angluin and Laird, 1988]. Direct or indirect examples may be given. Direct examples would be explicit at each stage the machine needed to make a choice. With indirect examples the agent or algorithm must additionally determine the value of any given decision, rather than it being pre-assigned. The program is (for instance) given a final goal, yet the value of particular intermediate decisions is not provided. Indirect examples make the work of the machine more complicated as it must de-

termine which of the intermediary decisions it made were most responsible for the failure or success of the trial. This is known as credit assignment. In general, all the examples available are presented to the algorithm in an order chosen by the creator/teacher, or a random order, but in some types of learning the program itself can ask for examples with specific characteristics missing from its current understanding to help it build a model. Active learning, described in Section 2.2.3, is an example of this. Other algorithms, for certain applications, can act entirely without provided examples either by generating such examples itself (as in learning to play a board game by playing against itself) or by collecting its own examples from the environment, as might be the case with a robot learning to walk or drive in varied terrain.

As with any sample, the training examples will be more useful if they adequately represent the real population. Indeed, several methods exist for determining the appropriate sample sizes for particular types of classifier [Fukunaga and Hayes, 1989, Kalayeh and Landgrebe, 1983]. Ideally a sample set should have a similar central tendency (e.g. the same mean) and a similar spread of data (e.g. the same variance or standard deviation) as the whole population. Special care should be taken if the machine is providing its own examples as this could be a very narrow selection without the same coverage as the true data. A parallel from the discipline of computer security is “Schneier’s Law”:

“Anyone, from the most clueless amateur to the best cryptographer, can create an algorithm that he himself can’t break.” - Bruce Schneier [Schneier, 2011].

In analogy, for an example chess-playing machine: “Any machine that can play chess can do so well enough that it always draws with itself.” Finding appropriate training examples can be challenging depending upon the area of interest. Often in practice the distribution of the training set differs from the test set or the real-world situation.

2.2.1 Supervised Learning

In **supervised learning**, training data with **inputs** and **outputs** are presented, and the algorithm is required to return a function fitting the inputs and outputs it has seen. ML shares many similarities with statistics research, although the latter often uses a different nomenclature. In statistical terminology inputs and outputs are known as **predictors** and **responses**, or **independent** and **dependent** variables [Hastie et al., 2001]. Given a **training set** of pairs (x_i, y_i) a mapping from x to y is to be learnt, where y contains a label y_i for each x_i . The mapping is evaluated based upon how well it predicts the label y for particular examples [Chapelle et al., 2006]. The outputs might be a class for the **classification** task (discrete, non-ordered answers) or by **regression** provide a continuous answer sometimes following a linear function $Y = mx + c$, where $Y \in \mathbb{R}$). The discovered function is

then either validated against a separate **test set** of known examples or is **cross-validated**: all known examples are split into n sets, $n - 1$ sets form the training set, and one set is used as the test set. This is repeated n times with each set in turn performing the role of test set. **Leave-one-out** cross-validation is a specific type of cross-validation which uses all-but-one of the examples as the training set and one example to be predicted as the test set, a process then repeated such that each example takes the role of test set exactly once. Good performance achieved via cross-validation is seen as evidence against **overfitting** on the training data. An overfitted model is one which performs well on training examples, but has become too specialised to those specific examples and as such performs much worse when given unseen examples from a more general distribution.

Supervised learning is appropriate when many **labelled** (classified) examples are available [Kotsiantis et al., 2006]. There are two families of supervised learning methods: **generative**, and **discriminative**. A generative method models the distribution such that new examples can be generated, whereas a discriminative model merely performs classification and regression based upon labelled examples with no scope for generating new examples. Discriminative models do not translate easily to **unsupervised learning**, whereas generative models may. Examples of discriminative models include **Support Vector Machines**, **linear regression**, and **Neural Networks**. Examples of generative models include Gaussian (and other) **mixture models** and **Naive Bayes** [Vapnik, 1998].

2.2.2 Unsupervised Learning

In unsupervised learning, unlike in supervised learning, the learning algorithm is given only unlabelled examples. This removes the problem of having to pre-define the classes of examples, with the corollary that the output function is not labelled, but can only show the organisation of examples in the sense that groupings of examples have **similarity**. $X = (x_1 \dots x_n)$ is a set of n examples from a distribution X , with the assumption that the set of examples is independent and identically distributed with respect to X . The assumption of independence means that the values of the examples are not influenced by the values of any other examples, i.e. no examples are interdependent. Identically distributed means the samples are drawn from the same distribution, but moreover the sample set has the same mean and variance as X . Some methods can provide more explicit descriptions of what is similar about particular sets of data. Clustering algorithms and dimensional reduction such as principle components analysis or manifold learning are examples of unsupervised learning methods.

Principal Components Analysis (PCA) is a linear dimensionality reduction which represents a dataset with a number of vectors that is smaller than the number of dimensions. The vectors are the eigenvectors (inherently orthogonal) generated from the covariance matrix of the data. They are ordered such that the first ‘principal component’ is the vector which accounts for the most variance in the distribution, the second accounts for the next greatest

proportion, and so on, minimising the least squares error [Pearson, 1901]. Manifold learning makes the assumption that high-dimensional data lies on a lower-dimensional non-linear manifold within the high-dimensional space, and attempts to 'unfold' the data to that lower dimensional environment, like flattening out a map on a globe (or part thereof) to reduce it to two dimensions rather than three. This makes further data interpretation much simpler. Sammon mappings [Sammon, 1969] and Kohonen maps [Kohonen, 1982] are examples of manifold learning algorithms.

2.2.3 Active Learning

As mentioned in Section 2.2, some methods are able to work from a small set of labelled examples with a larger pool of unlabelled examples to request the most useful examples from the unlabelled pool to be given labels and returned as further input for the machine to analyse. The theory behind this idea is that the learner can consequently achieve greater accuracy with fewer examples. The requests it makes are termed **queries**, and the trainer or other information source providing the labelled examples is known as an **oracle** [Settles, 2009]. The provision of labelled examples is expensive in terms of oracle time, if a human expert, and in monetary cost or in computing cost in many cases. If the machine itself can determine the most useful examples, the costs of training can be reduced without reducing the performance of the machine.

2.2.4 Semi-supervised Learning

Semi-supervised learning exists in the space between supervised and unsupervised learning [Chapelle et al., 2006]. Similarly to active learning, semi-supervised learning operates on some labelled examples and many unlabelled: $X_L := (x_1, \dots, x_L)$ for which $Y_1 := (y_1, \dots, y_L)$ are provided, and $X_U := (x_{L+1}, \dots, x_n)$ without labels. Often this technique is used because of a high cost in obtaining labelled examples. Most approaches to this type of learning present the problem as unsupervised learning with additional information about the examples. A few approaches view it as unsupervised learning guided by constraints. In the case where some labels are never known, the latter perspective is more appropriate. There is no mechanism for the learner to request further labelled examples from the unlabelled pool, as there is in active learning.

Semi-supervised learning can be separated into **transductive** and **inductive** learning. Inductive learning outputs an overall prediction function for the entire space of X , whereas transductive is focused on specific test points for prediction without the overarching function. Transductive learning can be thought of as X_u being the test set and X_l being the training set [Arnold et al., 2007]. The applicability of semi-supervised learning depends on the ability of the unsupervised examples to add information that is relevant to the classes or regression function as found in the supervised examples. One such assumption is that if x_1 and x_2 are close, y_1 and y_2 are also close.

Without the assumption that close parameters in feature space imply close outputs, it would be as impossible to generalise from the data. Necessary assumptions such as this are known as **inductive bias** [Mitchell, 1997]. More generally, the inductive bias of any learning system consists of the assumptions it uses to generalise from the input data to unseen examples. Every learning system has an inductive bias. Two common sources of bias in existing learning systems are (1) the generalization language is not capable of expressing all possible classes of instances, and (2) the generalization procedure that searches through the space of expressible generalizations is itself biased [Mitchell, 1980].

2.2.5 Reinforcement Learning

Autonomous agents use reinforcement learning to improve their responses to the environment. This is achieved by a reward/punishment system for appropriate or inappropriate actions. This can be inherent in the environment or administered by a trainer. For example, a robot might determine that it gets rewarded when it moves closer to, or perhaps simply reaches, a certain position, and punished when it makes contact with obstacles, and from this derive an obstacle-avoidance path-planning algorithm. Reinforcement learning is achieved by maximising the reward and minimising the punishment. The rewards and punishments can be direct or indirect: rewarding forward motion would be direct; rewarding the final state of having reached the goal

would be indirect since the intermediate states must be interpolated by the algorithm [Mitchell, 1997].

2.3 Learning Algorithms of Interest

2.3.1 Bayesian Reasoning

Arguably the basis for all rational learning, Bayesian reasoning is probabilistic inference: it allows calculation of the probability of a hypothesis being true given previous knowledge and new evidence. Thus, the probability of a particular hypothesis being true may be constantly updated as new information is encountered.

Bayes theorem:

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

To determine the probability of an event b given that event a has occurred, multiply the *a priori* (beforehand or initial) probability of event b by the probability of event a given that b has occurred, then divide by the *a priori* probability of a . The result is known as the *a posteriori* (after accounting for priors and performing computation) probability of b .

A Bayesian classifier returns the set of all hypotheses with its degree of belief in each one (the probability). A true Bayesian classifier is perceived

as the ‘gold standard’ by which other classifiers are judged, since it always outputs the ‘right’ answer (as a probability) for the inputs it is given. The main problem encountered is with encoding the prior knowledge accurately – most factors and probabilities of the real world are not known or at least are not precisely known. Bayesian classifiers are also computationally expensive. More often simplified versions are used for practical applications, such as the naive Bayes classifier, which assumes the conditional independence of all *effect* variables (conditional on the *cause* variable). The size of the representation of Naive Bayes models grow at $O(n)$ (linearly) with the number of effects, rather than the $O(n^2)$ of true Bayesian classifiers. Naive Bayes classifiers therefore can operate at scale, and perform surprisingly well even when the conditional independence assumption (that the probabilities of all effects are independent from one another) is wrong. Bayesian Networks are a different way to reduce the computational complexity by representing conditional independence in (directed acyclic) graph form [Pearl, 1985]. In a Bayesian Network each node is annotated with quantitative probability information and arranged such that the parents of nodes are those variables which have direct effect on the children. This effect is expressed as the conditional probability $P(X_i|Parents(X_i))$. Thus, the topology of the network defines the conditionality of the probabilities, and the values at each node give a conditional probability table. This graph could be created by a domain expert or, more usually for complex examples, generated from the data.

2.3.2 Decision Trees

For decision trees, inputs are described by attributes, which could be discrete or continuous. Each node in the tree is a decision based upon the value of a particular attribute, with possible values (or ranges) labelling the branches from that node. The “leaves” (from which there are no further decisions or “branches”) give a final classification or regression answer. The attributes are usually prioritised by which attribute’s value would give the greatest gain in knowledge, such that the root node is the attribute which has most effect on the outcome [Navada et al., 2011]. Decision trees are popular inductive inference algorithms. Their inductive bias is a preference for small trees over large trees [Mitchell, 1997]. This means that where all else is equal, the simplest solution is chosen, that is, the tree with fewest nodes.

2.3.3 Artificial Neural Networks

Artificial Neural Networks (ANN) are a non-symbolic approach to learning, inspired by the current understanding of how real biological neurones operate. Networks of artificial neurons ‘fire’ (transfer a signal to a forward node) given summed input values above a threshold from the neurons behind them. Each neuron holds a weighted sum of its inputs determined from the weights associated with each link, which may be positive (increase the sum in the receiving node) or negative (decrease it). There are several problems with

neural networks. One is the amount of space and memory they take up computationally, which has in the past been problematic. The other problem is insurmountable for certain applications: that a neural network is essentially a black box. No rules about how it has come to its conclusion can be discovered, at least, not simply [Benitez et al., 1997]. The first ANN was a single layer perceptron, the simplest possible ANN with all the inputs connected directly to the outputs. Since for a threshold perception the weighted sum of the inputs is equal to $\sum_{j=0}^n W_j x_j > 0$, this defines a hyperplane in hypothesis space. The threshold perceptron can be termed a linear separator. A linear separator can represent Boolean AND, OR, NOT, and further more complex functions like 'more 0s than 1s' very compactly, but it cannot separate functions like XOR because there is no single straight line that can be drawn between the two classes. Adding a hidden layer between input and output overcomes this problem by allowing functions other than straight lines, flat planes and hyperplanes to be defined. Hidden units can be combined to create the desired function, though it is by no means trivial to determine how many hidden units are needed for a given problem. Multiple layers (i.e. at least one hidden layer) are almost universally used now when a neural network is applied. This extends to the newer development of deep belief networks, which are probabilistic generative models comprising several simpler learning modules, each of which is a two-layered Boltzmann machine with one layer for representing data and one for learning higher-order correlations [Hinton et al., 2006, Hinton, 2009].

2.3.4 Support Vector Machines

One of the more recent and popular developments is that of the Support Vector Machine (SVM).

The basic mechanism of an SVM is to describe a hyperplane (in the two-dimensional case, this is simply a line, but planes or hyperplanes are required for operation in higher dimensions) which divides two sets of data in multidimensional space, by finding the hyperplane that is furthest away from both classes, a maximum-margin hyperplane. The support vectors are those samples on the margin delineating each class from the others [Cortes and Vapnik, 1995]. Although the original SVM is a linear classifier (a linear separator like the single-layer perceptron), by translating a non-linear problem (such as the XOR example given in Section 2.3.3) into a higher dimension using some kernel, solving with linear SVM, and then translating back down again, the linear SVM can be used to solve non-linear problems. This is known as the “kernel trick”, a technique which has been applied to several different linear classifiers. The idea behind a kernel function is to embed the data in a space where the resulting pattern can be described linearly. A simple example might find the straight line $y = mx + c$ of best fit given data that has been translated into a feature space such that the examples appear in a pattern which is easily separated by a straight line. The resulting function can be transposed back into the original feature space [Shawe-Taylor and Cristianini, 2004].

Support Vector Machines have been shown to perform very well in classification tasks such as genre identification, and are often applied in combination with other types of classifier giving increased rates of success [Li and Ogi-hara, 2003, Eisenthal et al., 2006]. See Chapter 3 for applications of SVM to music classification.

2.3.5 Inductive Logic Programming

Inductive Logic is a meeting of machine learning and logic programming. It operates on the premise that the combination of background knowledge and hypotheses entail (lead to) the examples. Examples, background knowledge, hypotheses and classifications are all stored as predicate logic. It is usually applied to domains such as natural language processing and bioinformatics because of its expressive power and its ability to utilise background knowledge [Muggleton and de Raedt, 1994]. It has also been applied to other domains such as engineering and market research. More recently attempts to combine logic, learning, and probability theory have resulted in the new sub-field of Probabilistic Inductive Logic Programming, also known as Statistical Relational Learning.

2.3.6 k -Nearest-Neighbour algorithms

The k -Nearest-Neighbour is one of the simplest machine learning algorithms. In a k -Nearest-Neighbour or kNN, test examples are classified based upon the classes of k (a number chosen by the user) labelled examples which are closest to the test example in feature space (an imaginary region in which each feature represents a dimension). The value of k generally needs to be higher the more noise there is in the dataset – more example neighbours are needed to give the right classification when those examples are noisy. Determining an appropriate value of k is important for getting the best learning performance. Often the feature space and distance measure is considered to be Euclidian, but not always. Inputs must be numeric and so for categorical data a suitable translation must be performed or alternatively a non-continuous distance metric is used such as Hamming distance (the number of features which have a different value). The algorithm is:

- Given a query example e , find the k examples which are closest in featurespace to e
 - EITHER: Using a majority vote, assign e the value most represented by the y (output) values in the k examples. (this means k should usually be odd-numbered)
 - OR: for regression, take the average of the y value for all k examples.

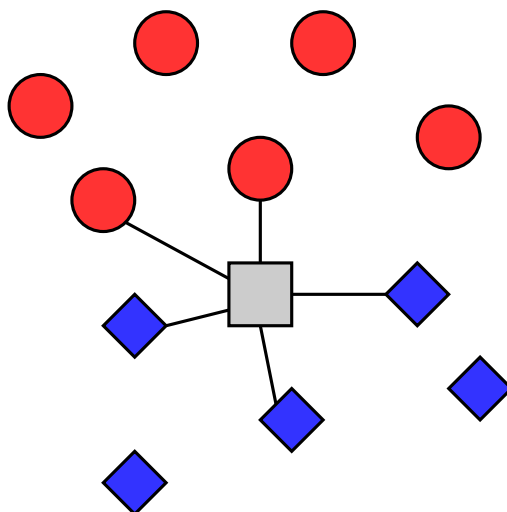


Figure 2.1: kNN diagram. In this case, $k = 5$, and the classification is blue (diamond) at a 3 to 2 vote.

Some kNN algorithms use weighting. The simplest of these inversely weights the influence of each labelled example with the distance of the example from e [Dudani, 1976]. This is especially helpful when a large proportion of the examples have a particular value so as to reduce this bias in the result.

2.4 Machine Learning Tools

Weka is a popular data mining suite written in Java which offers many different machine learning algorithms and parameter options. It is open source, freely available and has an accompanying book Data Mining: Practical Machine Learning Tools and Techniques [Hall et al., 2009] [Witten and Frank, 2005]. Being written in Java means that it is platform independent and

fairly simple for a developer to extend with new algorithms. It also provides several helpful interfaces to running and analysing experiments as well as visually exploring data for patterns. The usual data format for Weka is the Attribute Relationship File Format (ARFF) file, though Comma Separated Value (CSV) files may also be used for input. An example file has two parts: a description of each attribute (or feature) and the data itself as rows.

2.4.1 Example ARFF file

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
% (a) Creator: R.A. Fisher
% (b) Donor: Michael Marshall (MARSHALL@PLU@io.arc.nasa.gov)
% (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Weka's available learning algorithms fall into various different categories, which are:

- bayes: includes BayesNet, Bayesian Logistic Regression, NaiveBayes
- functions: includes Logistic, Multilayer Perceptron, Sequential Minimal Optimisation of SVM
- lazy: includes IB1, IBk, KStar
- meta: includes Ensemble of Nested Dichotomies, Bagging, Random Committee
- mi: includes multi-instance classifiers
- rules: includes ZeroR, OneR, decision table
- trees: including J48, RandomForest, REPTree
- misc: anything that doesn't fit into other categories e.g. HyperPipes

This is not an attempt to provide an exhaustive list as this changes with each new version of Weka available. It is unnecessary to go into depth about each specific classifier, but those which prove successful will be more closely examined. Some algorithms naturally operate only on certain types of data and can be immediately discounted from experiments where they are not applicable. Indeed, Weka itself highlights those algorithms which are feasible for the task using the ARFF file information. This does not necessarily mean they will perform well, but simply that the data format is appropriate for

the algorithm in question.

Other examples of machine learning suites include RapidMiner [Mierswa et al., 2006] which is a framework fully integrating Weka, and ELKI [Achtert et al., 2008] which focuses on unsupervised learning methods.

Chapter 3

Machine Learning, Aesthetics and Musical Aesthetics

3.1 Machine Learning and Aesthetics

Machine learning techniques have been applied to the aesthetics of other forms of art and to natural aesthetics with success. Eisenthal *et al.* looked at facial attractiveness in 2006 [Eisenthal et al., 2006]. An aspect that was once believed to be subjective (in the eye of the beholder), it has been shown that people generally agree on what is an attractive face, irrespective of ethnicity, social class, age and sex. Eisenthal set out to train a machine learning algorithm (in this case, k-nearest-neighbours and Support Vector Machines) to distinguish between attractive and unattractive faces. Human

participants first rated a set of images of people's faces. The middle 50% of faces as rated by human participants was removed from the sample such that only the bottom 25% and top 25% remained. The kNN algorithm performed best achieving 77% correct on the second data set tested. Since kNN by its nature could not provide any insight into the rules behind the ratings, a decision tree was employed to determine the most important features. The best were: size of the lower part of the face (jaw length, chin length), smoothness of skin, lip fullness, and eye size, all of which were found consistent with previous psychophysics studies. Averageness and symmetry are key parts of facial attractiveness across cultures [Rhodes et al., 2001]. However, there is more to facial attractiveness than these aspects alone [DeBruine et al., 2007]. DeBruine found that there are specific nonaverage characteristics that are particularly attractive, by using manipulated images on a scale. The most average face in each set was not rated as highly as some nonaverage faces with exaggerated features such as thinner eyebrows, bigger eyes, smaller chin. When these features were so exaggerated that they fell outside normal bounds, however, they were rated less well, suggesting that averageness is an important component but not the only one.

One might be tempted to suggest that this sort of analysis is only successful because facial attractiveness is such a close biological and evolutionary imperative in humans, or perhaps that natural feature aesthetics are a different problem from created art aesthetics. However, this appreciation extends to birds, fish and automobiles [Halberstadt and Rhodes, 2003]. Halberstadt

and Rhodes manipulated images of birds, fish and automobiles for averageness. Both manipulated averageness and rated averageness were strongly correlated with attractiveness for all three categories. However, when the effect of subjective familiarity was partialled out of the test only the bird and fish categories retained significance. This suggests that our perception of beauty in living things is indeed different from how we perceive inanimate objects.

Beyond living beings and inanimate objects, there are more general applications of aesthetic prediction. Datta *et al.* used general photographs as a basis for aesthetic analysis [Datta et al., 2006]. In the study 56 low-level features for the 3581 images were chosen based upon known aspects relevant to photography. Using information about lightness, colourfulness, hue, saturation, the ‘rule of thirds’ from photography theory and various other features for each photo, they applied SVM and CART algorithms to classify photos into the high or low category as well as using linear regression to produce a point on a scale. The best results for binary classification were 68.08% accuracy for predicting high aesthetics, and 72.31% for predicting low aesthetics when applying 5-fold cross-validation and using SVM. For the regression task the residual sum-of-squares $R_{res}^2 = 0.5020$, a 28% reduction from the variance $\sigma^2 = 0.69$. The generated decision tree produced interesting paths, one notably that a picture is highly rated when it is closely focused on a central object of interest. The human ratings for the photographs were obtained from Photo.net, a social networking photo upload site, where

users upload photographs they have taken and rate other photographs on a scale [Photo.net, 2013]. Ideally the authors suggest a controlled study for a more balanced demographic, but stated a lack of time as justification for using the abundant and freely available information on Photo.net.

In general, some progress has been made in tackling the problem of aesthetics in imagery in the last decade [Joshi et al., 2011].

It is clear from the above studies that different kinds of visual aesthetics can be modelled to some extent with machine learning techniques, despite their apparent subjectivity.

3.2 Musical Aesthetics

Eduard Hanslick, writing in the Romantic era, fought against contemporary ideas that music was for stirring the emotions [Hanslick, 1891]. He tried to establish that the aesthetic value of music could be understood purely intellectually. He insisted there were means by which to judge music that were not purely subjective and whose basis did not lie in how the music made people feel. Nearly 200 years later there is still debate over what is important in the judgement of music. Despite (or perhaps because of) this, work has been conducted into measuring the aesthetic value of music, with the idea that aesthetics are valuable in music, and some agreement can be formed on their nature, just as in other forms of art (e.g. paintings), despite the

difference in medium. Roger Sessions writing in 1970 came to very similar conclusions to Hanslick regarding the criteria by which to judge music being inherent in the music, and not relating to emotions or celebrity, or the process by which the music was created [Sessions, 1970].

In 2003 Manaris *et al.* conducted a study using a statistical approach to aesthetics [Manaris et al., 2003]. As explained in Manaris’s work, Zipf previously analysed celebrated works of literature for their word usage distribution, and found that the frequency plotted against the logarithm of their statistical rank produces a line with gradient close to -1. Put another way, the frequency of the n^{th} ranked word is $1/n^a$ where a is close to 1. Benoit Mandelbrot extended this law to account for a wider range of phenomena, where the slope of the lines range between 0 (random) and negative infinity (monotonous).

The study by Manaris postulates that music that is aesthetically pleasing should follow a Zipf-Mandelbrot law, which is a type of power law. They tested several different aspects of music against this law, using MIDI (Musical Instrument Digital Interface) examples of music so that calculations could be easily automated. The aspects covered pitch, duration, melody and harmony in various ways, including a pitch mod-12 aspect which would likely only apply to Western music because of the strict implication of a Western scale here.

Celebrated classical musical works were compared with 12-tone, pop music,

jazz, white noise, pink noise and music generated from DNA. The Zipf law was in general present in musical works and not present in non-musical noises. A neural network was applied to distinguish between 100 pieces by Bach and 32 pieces by Beethoven using these Zipf metrics as features. 95% accuracy was achieved on unseen data (unfortunately, the authors did not report the proportions of training and test data). The research concluded that although a Zipf-like distribution is a good start to a musical piece it does not on its own guarantee it to be aesthetically pleasing. What can be said is that if a sound has a distribution far from Zipf in several dimensions, it is either not music, or not aesthetically pleasing music.

The use of MIDI also leaves a lot of information out of the calculations, but reduces the complexity of the task significantly. MIDI encoding is very capable of representing polyphonic electronic music, but vocals cannot be encoded, nor can some aspects of musical expression such as gradual changes in timbre. Timbre is the quality of the sound which is additional to the pitch (frequency) and loudness (amplitude). Timbre is what creates the difference between playing middle C (440Hz) on a piano and on a violin at the same volume. Overtones and harmonics contribute to timbre. Actually, MIDI contains no information about how the music sounds in terms of timbre - only information about the note pitch, length, and suggested instrument (and a few other things too such as volume and pitch bend), which a soundcard can interpret to produce 'music'.

In 2001 *Music Cognition, Culture, and Evolution* by Ian Cross was published [Cross, 2001]. At first music seems irrevocably cultural and thus a scientific generalisation is useless. However, all music has this in common: temporal organisation which is regular and periodic. In fact several aspects of music appear to be cultural universals. Dowling and Harwood in “Music Cognition” [Dowling and Harwood, 1986] identified the following cross-cultural universals:

- use of octaves
- a logarithmic and discrete pitch scale, with 5 to 7 pitches per octave
- hierarchical structure
- melodic contour (seen as a series of curves in written music)
- timbre changes
- the existence of a beat framework
- rhythmic contours
- every culture has people who can sing

The above list indicates that whilst some aspects of music comprehension may be understood only from within the culture in which the music was created, many other aspects appear to be universal to humanity at large. Levitin writing in 2006, a neuroscientist who previously worked in a recording studio, has discovered many of the same universals [Levitin, 2006]. This is

encouraging for the planned cross-cultural nature of this thesis.

3.3 Machine Learning and Music

The work that encompasses music and machine learning invokes both “sound to sense” (interpreting music), and “sense to sound” (creating music) paradigms, as well as including associated analysis and translation tools. These terms come from the sound and music computing network, <http://smcnetwork.org>. This part of the literature review concentrates on those advances in “sound to sense”, since we are interested in the interpretation, rather than creation, of music by computers. Despite the quiet growth of the field of sound and music computing, and despite the lucrative potential of such endeavors, academic support for such work is not yet mature [Consortium, 2007]. The S2S2 roadmap, a projection of the direction of music information retrieval research and identification of stumbling blocks and useful new research that should be conducted, expresses the need for music perception models, since “bottom-up” modelling (that which deals with the raw sound features and does not extend to modelling or mimicking human responses) has reached its limitations.

A large amount of research has been recently done on the development of audio features (attributes) for the computational analysis of music, e.g. [Grachten et al., 2009]. These attributes have typically been used to perform

automatic classification and clustering to identify similar pieces of music (for recommendation systems), e.g. to identify mood, genre, emotive content, and various other purposes for which it would be impossible to provide an exhaustive list [Laurier et al., 2008]. In addition to audio attributes other meta-data have been utilised via, for example, web searches and social tags, but also MIDI, score reading and lyric mining [Widmer et al., 2005, Turnbull et al., 2009b, Knees et al., 2007, McKay and Fujinaga, 2004]. Features extracted from music are discussed in Section 3.3.1.

Various machine learning methods have been applied to such audio features. Support Vector Machines, k -Nearest-Neighbour, and Neural Networks have all been used with good success for certain applications. The MIREX (Music Information Retrieval Value Exchange) 2012 competition winners for classical composer identification used a SVM ranking algorithm which achieved an overall classification accuracy of 69.7% on 11 classical composers. The mood classification task was also won by a team using SVM extensively, though their winning algorithm also used Support Vector Regression and hierarchical clustering. The genre recognition task was also won by a team using SVM, whose algorithm achieved 76.1% overall accuracy discriminating 10 genres. Though other algorithms are in use and do well, SVM is clearly the go-to machine learning algorithm for state of the art music classification. Hamel and Eck used Deep Belief Networks (DBN), a newer form of neural network, to generate features which were then used as inputs to an SVM classifier [Hamel and Eck, 2010]. The results compared favourably with the state of the art

for both genre classification and tagging tasks. k -Nearest-Neighbour algorithms are more often used for symbolic tasks [Rebelo et al., 2010]. Rebelo *et al* compared Hidden Markov Models (HMMs), k -Nearest-Neighbour, Neural Networks and SVMs in several symbolic recognition tasks on handwritten and printed music. The simple kNN performed better than either HMM or NN, beaten only by the SVM. Accuracy ratings varied significantly depending on the symbol being recognised, for example the handwritten open note was extremely difficult to classify, with even the SVM getting accuracy results in the 40% region, to sharps and naturals for which all the algorithms performed in the 80-100% region whether handwritten or printed.

Despite these advances little work has been done on ‘Computational Ethnomusicology’.

Tzanetakis’ work [Tzanetakis et al., 2007] seeks to illustrate the usefulness of ‘Computational Ethnomusicology’, which means the application of music information retrieval (MIR) techniques to ethnomusicology; that is, to musicology not exclusively focused on Western (and usually classical) music. Historically most music analysis work has been on Western music. MIR allows the analysis of large corpuses of music to obtain automatically features that would take copious time to transcribe by hand. Though the features are from signal processing they relate closely to how humans perceive music. One example is the spectral centroid, which is mathematically simple (it is the weighted mean of the frequencies in the signal) and yet is strongly cor-

related with human perception of ‘brightness’ in sound [Lichte, 1941, Grey, 1977].

Few examples have been offered, such as Liu *et al.* which demonstrated the applicability of music analysis techniques to non-Western music [Liu et al., 2009]. They took 1300 tracks matching 6 different cultural styles: Western classical, Chinese traditional, Japanese traditional, Indian classical, Arabic folk and African folk music. SVM, kNN and decision trees were trained on mainly low-level features of the sound. The features monitored were several timbral, rhythmic, wavelet and some higher-level musicology features (these last including chroma features which represent the distribution of the 12 pitches in the Western scale). Of these, the timbral features, important for distinguishing instruments, were found most useful, performing with 84.05% accuracy alone. One might question the applicability of Western features to the non-Western styles, since the notes of some cultures would not line up with the pre-defined Western pitches. However, the musicology features did increase performance, albeit by 1-2 %. The rhythm and wavelet features did not improve performance for this task. This may be, as posited, because these aspects do not contain information relating to cultural style, but it could be that the features were not detailed enough to pick up on such differences, too. Some classes were better classified than others, ranging from 97.33% for Western classical down to 68.57% for African folk. As the authors suggest, this is probably due to greater diversity in style within Indian, Arabic and African music.

Gomez *et al.* have applied these techniques to classifying music as Western or non-Western with success, and also found some important features relating to the latitude and longitude of origin of a piece [Gomez and Herrera, 2008, Gomez et al., 2009]. In their 2008 paper they attempt to distinguish Western music from non-Western music, with 1000 and 500 examples respectively. Assuming that the first 30 seconds of each track is enough based upon previous work with key estimation, they extract several tonal features including deviation from western tuning frequency (440Hz), high resolution pitch class distributions at 10 values per semitone, much more finely grained than the 12 per octave used by Liu *et al.*, a transposed version of the same which is invariant with absolute pitch, and a measure of roughness (close frequencies played together). Several machine learning techniques were applied and tested using 10-fold cross-validation. Their decision tree and SVM learning choices are described, but all their tested algorithms performed at above 80% accuracy, with the best-performing being SVM at 86.51%. Cluster analysis using k-means set to find 2 clusters showed a close correlation with the found clusters and the actual categories. This work demonstrates the applicability of tonal features to the task of distinguishing Western and non-Western music.

In Gomez *et al.*'s 2009 paper the aim was to firstly distinguish Western and non-Western music for a larger dataset than the previous paper, and secondly identify those audio features which distinguish music geographically. For this tonal, rhythmic and timbral features were selected. The tonal features are

as in the 2008 paper, timbral features are a standard set including spectral flux and roughness, and rhythmic features include global tempo and onset rate, which was also broken down to drum-kit events such as onsets of bass drum/hi-hat (whether the real instrument playing the detected sound was actually this type of drum or not). An SVM was trained on these features to classify tracks based on the features from the first 30 seconds of each. The classifier reached its highest accuracy – 88.53% – with timbral features alone. As before, rhythmic features performed badly, as did the drum-based features. They then computed the Pearson correlation coefficient for latitude and longitude. Latitude was correlated with mostly tonal features, and longitude more with rhythmic features. There are some problems with this approach, however. One is that the zero line for longitude is arbitrary and so predicting on a differently centred longitude line might give different results. In flattening the map such that -180 degrees is as far as possible from +180 degrees when on a globe they are along the same line fails to account for true distances between locations on Earth. The combinational contribution to latitude and longitude is not addressed, they are treated as orthogonal and independent, when cultures that are close are likely to be close both in latitude and longitude. It is these potential areas for expansion which inspired the work of Chapter 5.

3.3.1 Source Material and Representations

Signal and Audio

Source material for interpretation takes many forms. One obvious source is the music itself. There are good arguments for using data closest to its raw form as possible, but any kind of music must necessarily be digitally converted for a machine to interpret it, which means the interpretation deviates from human perception. The difference between being at a live concert and listening to a CD is obvious, but so far we can only really provide the auditory information, possibly with some data-mined context in textual form, and recently, the actual motions of the performer have been encoded and included to help with interpretation. The format of the music has generally been whatever is readily or freely available, such that most recent work is applied to MP3 files.

Earlier work was mostly conducted on MIDI files, which have already encoded structural aspects of the music, and by no means represent the entire auditory content. These were used because the representation was to an extent already there, and because the amount of data was very small in comparison to an entire track of auditory data. More recent developments in processing power and storage capability have led researchers to use audio data as far as possible. However, MP3s are optimised for human ears, and as such the loss of data might have somewhat less effect on the ability of the computer

algorithm to emulate human listening capabilities, particularly when they are currently behind humans at the most complex recognition tasks.

Many of the low-level audio features are based on the Short-Time Fourier Transform (STFT). Fourier transformations are a way of expressing a signal that is expressed as a function of time into a frequency spectrum. The nature of digital signals requires such transforms to be Discrete Fourier Transforms (DFTs). Music signals change over time. To incorporate this, rather than taking a DFT of the whole track, one may take STFTs of shorter overlapping windows and take the DFT of each of these. The size of the window affects the accuracy of both frequency estimation and time resolution, such that a larger window gives better high-frequency resolution, but a smaller one gives better time accuracy. This trade-off is fundamental to time-frequency analysis.

Filterbanks are systems which separate input into several sub-bands. In a sense, even an STFT can be considered a filterbank, along with wavelets and other signal decompositions. The Mel filterbank is based upon the Mel scale, which approximates the way human ears perceive the loudness of different frequencies.

The dominant feature that is extracted by MIR specialists from raw audio data files is the Mel-Frequency Cepstral Coefficient (MFCC). This is a means of summarising Fourier transforms over a window of time, such that they follow a logarithmic scale that mimics the way in which the human ear

perceives sound. The scale which mimics human perception is called the Mel-Frequency scale. MFCCs have been shown to perform well with not only raw audio data, but also MP3 files, so long as the bitrate (data density or data quality, essentially, as measured by the number of bits per second recorded) is at least 128kbits/s [Sigurdsson et al., 2006]. Timbral features like the MFCC have been the most widely used and the best-performing features in isolation.

Another common type of timbral feature are spectral features. These are summaries of the general distribution of energy across the frequency spectrum. Spectral centroid – a measure of ‘brightness’ in sound, and spectral rolloff is defined as the frequency below which 85% of the energy distribution of of the spectrum is concentrated.

Other timbral features include zero-crossing rate, spectral bandwidth, octave based spectral contrast, spectral flatness measure, and spectral crest factor. Timbral features can be thought of as summaries of the frequency information. Temporal summation of timbral features is often useful for giving a single feature vector per song or track.

Rhythm features include representing the tempo or, with more difficulty, the beat (which may have nuances like strong/weak and syncopation) of the music. Usually audio onset detection is used to translate the audio representation into a symbolic one, and then applying symbolic algorithms.

Pitch or harmony features extract the notes and chords from audio. A com-

mon way to extract pitch or the ‘note’ is the Pitch Class Profile, which folds the pitches extracted into one octave representation. One approach to this computation is known as the chroma or chromagram. Each Fast-Fourier Transform (FFT) bin is mapped to its closest note, usually on a Western scale of 12 notes, segmenting the spectrum into note regions. The result is a vector of size 12 with each value corresponding to the magnitude of that group of frequencies.

Many of these features have a short explanation in Table 5.1.

Written and Symbolic

A second source often used is a written, rather than auditory, logical representation of the music, such as a musical score. This is known as ‘symbolic data mining in musicology’. Score analysis is mostly used on classical Western music, since non-Western music is often not written in the same way or even at all, and non-classical music tends to have just the chords and basic riffs without the entire music being represented. A recent useful application of score analysis was to determine the difficulty of pieces such that they could be sorted by ease of playing for students [Sébastien et al., 2012]. The music score in this case was in MusicXML format. The proposed criteria to represent difficulty were playing speed (using tempo and the length of the shortest note), fingering (how awkward the hand positions are, applying a cost function to the notated positions), hand displacement (how far

apart the hands are, adapted per instrument), polyphony (number of notes played simultaneously), harmony (number of accidentals [notes outside the expected scale often introduced for interest or flair), rhythm (synchronising rhythms of difficult ratios such as 3 against 2), and length (the total length of the score). They were able to extract most of these criteria (not fingering as their work to extract this is not yet complete) as features from 50 piano pieces ranging in difficulty from beginner to virtuoso with the bulk in the intermediate to advanced range. They used k-means clustering on these features and compared their results with PCA and with expert ratings (from music teachers). There was a better correspondence with the human ratings than with the ‘objective’ PCA comparison, reinforcing the validity of the chosen criteria.

In 2008, Inductive Logic Programming was applied to musical harmony by Amelie Anglade [Anglade and Dixon, 2008]. The chord data was manually annotated in Resource Description Framework (RDF) format, rather than automatically extracted. The chord, degree of the scale, and intervals between chords were recorded. Over twelve thousand underlying harmony rules were extracted, the most useful of which could distinguish between and give the characteristics of a set of music by The Beatles, and jazz music. Known musical harmony rules were amongst those discovered by the algorithm.

Niedermayer and Widmer analysed the influence of having real audio vs. basic synthesis from MIDI or better synthesis from a more expensive program

(the real instrument, a Boesendorfer SE290 piano, was computer controlled). Real performances of 13 Mozart Sonatas were performed on the SE290, and the result was matched note for note with a written score. Algorithms known to perform well from e.g. MIREX competitions were selected for the tasks. For onset detection (finding the onset of a particular note) the quality synthesiser performed best with mean f-measure 98.18, performing better than the ostensibly simpler basic synthesiser (94.00). This was surmised to be due to the influence of higher frequencies in the basic synth. For audio alignment (the task of matching audio to score) the real piano produced the best result of f-measure 86.85. Lastly they investigated the influence of expressivity on these two tasks by ‘cleaning’ the MIDI to set it to standard timings rather than those actually played. They found that contrary to their expectations, onset detection was not helped significantly by aligning the notes to their score positions, but audio alignment was made easier by having the expressivity removed. The conclusion was that whilst synthesised music can be useful for mining musical information it can also be prone to risks of overfitting.

Metadata and Tags

A third kind of source material is semantic data from the social context, almost always from the web. Types of data mined include song lyrics [Knees et al., 2005] [Geleijnse and Korst, 2006], country of origin [Govaerts and

Duval, 2009], band members and the instruments each plays [Schedl et al., 2007], and tags for genre, mood and other aspects [Bertin-Mahieux et al., 2008]. These are found on web pages about the artists or tracks, lyrics sites, and tagging systems such as Last.fm. Access methods are scraping web pages, interacting with APIs such as that of Last.fm, eliciting participants to play tagging games [Mandel and Ellis, 2008] or manual searching.

Recent work has attempted to combine different kinds of features with improved results than either of the groups could achieve alone [Turnbull et al., 2009a] Turnbull *et al.*. They considered two types of audio feature: MFCCs for timbre and chroma for harmonic content. Two social sources were used: social tags, and web documents.

The first 13 MFCCs are computed per time interval and the first and second instantaneous derivatives (deltas). Instead of summarising the entire song or a segment of it with MFCCs for the whole, these time intervals are about 23 milliseconds, resulting in 5000 39-dimensional features for 30 seconds of track. The full tracks were modelled using this with an 8 component GMM. Supervised labelling was applied to automatically allocate tags to the tracks, and these tags were similarly modelled for audio content using a 16 component GMM.

The social features were firstly tags from Last.fm, with a song having a relevance score for each tag based on a) whether the specific song is tagged with it, b) whether the artist is tagged with it, c) whether either is tagged

with any synonyms of the tag. Secondly web documents returned from a Google search of specific sites were used, with relevance weights for each document to the searched tag, and since each document was returned from a search querying a specific song, artist or album, a weighting for each song-tag combination is determined.

The goal of combining these measures is to create a ranked order of songs relevant to each tag. Calibrated score averaging which optimises the weight given to each ranking, RankBoost (similar to AdaBoost but designed for ranked data), and kernel combination with SVM (combining at an earlier stage to produce one ranking output) were each applied to combine the features. The performance of these combinatorial approaches was compared with performance on the separate feature groups and with a Single Source Oracle (SSO), designed to pick the single best source of a group based upon a test set. All the combinatorial approaches were better than all the single source ones at a 5% level of significance, but there was no statistical difference amongst the combinatorial approaches (including SSO).

They also calculated the number of tags for which each source was the best predictor. With direct ranking the MFCC best predicted 51 tags by direct ranking, web-mined tags 12, social tags 9, and chroma 0. For the SVM approach this was slightly more spread out at 42, 9, 21 and 0 respectively. It is notable that the chroma features, though perhaps contributing some information, were the most useful feature for precisely no tag, suggesting

these are at best an auxilliary predictor when it comes to music tagging.

McKay and Bainbridge created an expansion to the Greenstone open source library software package which extracts audio features and metadata from the stored tracks and keeps them in the repository to facilitate later mining by other researchers [McKay and Bainbridge, 2011]. This should make such research easier to conduct and thus more common in the future.

3.3.2 Audio Feature Extraction Tools

There are many available tools for extraction, including sox (a command line tool for audio manipulation found at <http://sox.sourceforge.net>), Audacity (a GUI-based visual tool for manipulating audio files found at <http://audacity.sourceforge.net>), JAudio (A Java suite for feature extraction) [McEnnis et al., 2005], LibXtract (a lightweight portable library of audio feature extraction functions) [Bullock, 2007], and MARSYAS [Tzanetakis and Cook, 2000]. MARSYAS was chosen for feature extraction as it is prominent in the MIR field and so gives the ability to compare our results easily. It also outputs in ARFF: a format suitable for Weka, though JAudio also does this. Many features are timbral in nature. The power spectrum is also commonly referred to - this is like the frequency spectrum except the amplitude is squared, so all values are positive.

MARSYAS Features

The MARSYAS features available are summarised in table 5.1.

3.3.3 Areas of Interest for Classification

Genre

The research question posed in most examples of genre research is “can the machine accurately predict the genre of given music after training with known examples?” Performance is at an acceptable level for most consumer purposes but is by no means perfect. In 2006 McKay published a paper in the proceedings of ISMIR 2006 that attempted to justify further work in the area despite the limitations and slowing improvements [McKay and Fujinaga, 2006]. Despite the difficulties associated with getting agreement about genre definitions, which change with time, the different words used by different cultures and countries, and the lack of objective ground truth, McKay and Fujinaga emphasise the difference between learning human-defined genres and learning musical similarity. Users are familiar with the notion of genre and expect to be able to specify it when searching for music. There are also relationships between genre and culture, especially for certain music genres such as rap or death metal. Social tags are used for recommendation extensively by online communities, and thus work is being done on the automatic generation of such tags [Bertin-Mahieux et al., 2008].

Feature	Explanation	Grouping
Time Zero-crossings	The number of times a signal changes sign, <i>i.e.</i> how often it crosses the horizontal zero line.	Default timbral features
Spectral Centroid	A measure of the “centre of mass” of the power spectrum.	Default timbral features
Spectral Rolloff	Describes the amount of skew in the power spectrum.	Default timbral features
Spectral Flux	Is an indicator of the amount of spectral variance based upon differences between adjacent spectral windows.	Default timbral features
Mel-Frequency Cepstral Coefficients	coefficients for a mel-frequency (tailored to human auditory response) power cepstrum - representing the short-term power spectrum.	Default timbral features
chroma	Detects frequency matches for each musical note of the Western scale (and its octaves)	Chromatic feature
Spectral Flatness Measure	Quantifies how tone-like, as opposed to noise-like, a sound is.	Non-default timbral feature
Spectral Crest Factor	Peak to average ratio of amplitude. Variance in loudness across frequency.	Non-default timbral feature
Line Spectral Pair	Describes the two resonance frequencies of the vocal tract when open or closed.	Non-default misc feature
Linear Prediction Cepstral Coefficients	As MFCC, but linear rather than Mel-scale	Non-default timbral feature

Table 3.1: MARSYAS features

The best feature for genre recognition is inarguably the MFCC [Fu et al., 2011]. Other low-level features are often combined to improve performance. Mid-level features like beat and pitch usually perform badly alone but can improve performance in combination with low-level features [Tzanetakis and Cook, 2002, Craft et al., 2007]. Craft *et al.* question the lack of research into industry and user classification of genre which is not the unquestionable ground truth that most approaches to genre recognition treat it as. They propose new evaluation criteria which takes into account the uncertainty in the result. It is their view that, in genre recognition, “ground truth is an artefact of an individuals response to music, not an artefact of the audio itself.”. Still, genre is a long-studied topic in MIR and in general research is spreading to other areas of music information retrieval (MIR) [Grachten et al., 2009].

The introduction to the special issue of the Journal of New Music Research - “From Genres to Tags: Music Information Retrieval in the Age of Social Tagging” discusses the progress of the field in the light of new uses of musical metadata [Jean-Julien Aucouturier, 2008]. Indications are that users are more interested in similarity than in absolute definitions for a particular track, and when such definitions are desired they are prevalent in a tag-cloud architecture such that a track may have many tags of varying types including genre, mood, style, instrumentation, and cultural associations such as films in which the song forms part of the soundtrack. Despite this user preference, a more rigid taxonomy may be of more use in determining underlying

rules for music.

Similarity

Similarity is a very popular topic within Music Information Retrieval, not least because of the commercial implications. Recommender systems, whether for music or for other applications, are legion. They operate in several different ways, often ignoring any musical information in favour of using a buying habits similarity measure such as that of Amazon's 'People who bought this also bought' system. Such systems are known as 'collaborative filtering' (CF). Systems which make use of content analysis are termed 'content-based filtering' (CB) [Park et al., 2012]. Park *et al.* categorised 210 papers from 2001 to 2010 into the data mining technique used and the application area. The keywords searched were 'Recommender system', 'Recommendation system', 'Personalization system', 'Collaborative filtering', and 'Contents filtering'. Of these, only 9 were applied to music, although they did not include conference papers because these were assumed not to be peer-reviewed. Actually for many computer science conferences submitted papers are rigorously peer reviewed, so it is unclear why they were not interested in these specific conferences. Another reason music recommendation may appear under-represented is the concentration on computer science journals rather than more diverse research areas. The application fields considered included books, movies, documents, TV programs, music and images, though the largest represented

group was the ‘other’ category. Publications relating to music recommendation do not appear in the dataset until 2007.

The 8 data mining groupings were:

- association rule
- clustering
- decision tree
- k-nearest neighbor
- link analysis
- neural network
- regression
- other heuristic methods.

In general, heuristic and kNN approaches were most popular. For music, clustering approaches were more prevalent, with 4 of the 9 papers using some clustering method for at least one experiment. With such a small dataset for music it would be unwise to conclude too much from this.

Interest in content-based similarity measures for music recommendation has increased in the last 10 years in MIR. Data is sparse in CF systems, and user bias, non-association, and cold start problems are aspects of working with sparse recommender datasets which can be alleviated by content-based retrieval. User bias is evident when preferred genres are taken into account

and this should be used in combination with the information about individual tracks. Tracks of the same genre could therefore be recommended before similar users have heard them. In a sparse dataset two items may never have been rated or wanted by the same user which tells us nothing about their relationship. With content information a relationship can nevertheless be inferred. The cold start problem refers to the need for a critical mass of ratings or likes before the system can function correctly. Content analysis can boost the number of ratings in the system allowing recommendation of new songs before any ratings are available. Li *et al* found that the addition of content information and genre information improved recommendations for a dataset with 760 16-bit MP3s [Li et al., 2007]. This is not good quality compression, so it is encouraging that despite the findings of Sigurdsson some useful information remains in lower bitrate MP3 encodings.

The use of timbre to determine the similarity of music has been said to have an upper limit [Pachet and Aucouturier, 2004]. Aucouturier and Pachet ran extensive tests on combinations of timbral features and learning algorithms but found that timbre features alone (including but not limited to MFCCs) with various learning algorithms being tuned as far as possible tends towards a limit 65% R-precision (a measure from text retrieval). They did not try other low-level features or SVMs, but suggested this for further work. They emphasised the need for new approaches rather than simply tweaking parameters and interchanging slightly different featuresets.

Magno in 2008 suggests humans are still slightly better at similarity than any other mechanism, but signal-based algorithms do nearly as well as services such as Pandora (human musicological analysis followed by automatic similarity algorithm) and Last.fm (collaborative filtering – if most of a user’s library intersects that of another user, the disjunction what-one-has-but-the-other-does-not should be liked by both users), suggesting that signal analysis is a useful tool for creating recommendation systems [Magno and Sable, 2008].

In 2007 methodological considerations were considered by Allan *et al.* in 2007. Music similarity can be in acoustic properties, or in cultural aspects. The relative importance of these aspects changes by listener and by context. In this study the focus was to represent user-expressed similarity. Users selected the 2 most similar pieces of three given. The algorithm is tasked to find further examples that are similar in the *same* way. Feedback training was used to further improve the results [Allan et al., 2007].

Popularity

There is some debate in this area over whether anything has been achieved in popularity studies, because there are so many confounding factors inherent in what sells well. The area is relatively new with surprisingly little attention from music psychology [Schellenberg et al., 2008]. Schellenberg *et al.* explored the effect of exposure to music on how much it was liked, and found

that – as one might expect – familiarity at first increases the liking of particular pieces and later, after too many listenings, decreases it. This 'inverted-U' shape was previously observed in a separate experiment by Szpunar, Schellenberg and Pliner [Szpunar et al., 2004]. Other studies, such as that of Witvliet and Vrana, show a different pattern of reinforcing and increasing initial responses [Witvliet and Vrana, 2007]. Witvliet also measures physiological responses and these reactions were correlated with reported ratings. It is clear that exposure affects ratings and experiential effects, but the precise way in which this occurs, especially across cultures, is not yet known. Since popularity affects exposure (a popular song gets more radio airplay, for example) and exposure affects popularity, disentangling this influence is no easy task. Salganik *et al.* studied social influence in preference. Two groups were created, one in which participants rated unknown songs independently and one in which they had access to other users ratings. The independent ratings were not uniformly distributed, suggesting that there is something about audio which affects whether people will like it. The ratings for the other group were more polarised than the independent ones, but also more unpredictable. The best songs rarely did poorly, and the worst rarely did well, but any other result was possible.

Dhanaraj and Logan [Dhanaraj and Logan, 2006] applied SVMs and boosting classifiers to lyric content and audio content to predict hit songs. They used probabilistic latent semantic analysis [Hofmann, 1999] to select topics for the lyrical content. An MFCC summary for each song was produced using

20-dimensional MFCCs and k-means clustering. One experiment achieved a success rate of 68% based upon lyric analysis alone, where the audio features achieved 66% accuracy. The found it difficult to improve further on these results, however. Whilst certainly better than random many aspects of the problem are difficult to take into account, such as those mentioned above. Pachet *et al.* [Pachet and Roy, 2008] contradicted this finding with a much more extensive study. They showed that the predictive power of audio and other features could be broken down to predicting subjective human labels affecting popularity, rather than popularity itself. 2 different acoustic featuresets and one of human-generated labels were used. The improvement over random classifiers in learning the popularity category was not significant, yet on other subjective categories such as mood it was significantly improved. Finally the first letter of the song was taken as a feature and classification improved by 5%, suggested by the authors to be an indication of noise (unless – which is possible – the letter with which a song title begins affects its popularity).

It is clearly an area in which work could be useful commercially, but it remains to be seen how plausible such work is given current and expected future technology. It is also questionable whether popularity can really be seen as a feature of a song, rather than its circumstances, at all.

Emotion and Mood

Emotion is considered to be how a piece makes one feel, and as such is often described as an ‘affective’ measure. The locus of ‘mood’ is more contained within the music itself than in the reaction of the listener - a piece can be said to have an ‘upbeat mood’ without the necessity of an upbeat listener. However, there is overlap between the two concepts and the ways in which they are studied for music. In a 2003 paper the detection of emotion with MARSYAS and SVM had a low success rate [Li and Ogihara, 2003]. This is in contrast to the same setup performing well on genre discrimination in the author’s Masters dissertation [Q, 2008]. The work did give the encouraging suggestion that within a similar cultural background, similar labels are chosen by humans. The low success indicates that old techniques may not apply to newer, vaguer areas such as emotion. However, some success with mood, rather than emotion, was achieved by discriminant analysis by Peeters, suggesting that either discriminant analysis is a better tool, or an externalised descriptor that is a property of the music rather than the person is an easier function by which to classify and generalise [G.Peeters, 2008].

Automatic mood estimation was considered by Skorownek in 2007 [Skowronek et al., 2007]. This is an area where personal taste is a factor, and yet ground truth was established, and cautious optimism was expressed about automatic mood estimation being achievable. Mood classes were defined where users mostly agreed on the moods which fitted the pieces. The example music was

chosen as the most easily classified music from a pool. Twelve final categories were created, on 1059 excerpts of music. The categories were not considered mutually exclusive, but ambiguous excerpts where most people could not clearly state definite belonging to a class or definite exclusion from a class were removed. This removed most of the excerpts leaving only the most unambiguous ones. Each piece was rated by 6 of 12 participants. Features used were:

- low level signal features
- tempo/rhythm
- chroma and key information
- percussive sounds

Quadratic discriminant analysis was used to distinguish between moods. They concluded that mood classification was possible when the ground-truth is strong and unambiguous.

In the *Psychology of Music* journal the topic of emotion and music appears with regularity. Konecni’s paper on music effects on emotional state vs. recalled life events’ effect on the same, the pattern of results and complex methodological issues cast considerable doubt on the idea of a direct causal link between music and emotion [Konecni et al., 2008]. It was also proposed that the notion of “musical emotions” be replaced by the concepts of “being moved” and “aesthetic awe”. Music here did generate weak emotions but

it is hard to separate this from potential associations or conditioning. Measurements were ratings of emotional feeling. The previous year, Thompson conducted a study on concert-goers who were attending performances of classical music [Thompson, 2007]. The survey found, among other things, that emotion predicts enjoyment much more than it predicts quality. From the above work and that on aesthetics it is apparent that emotional responses and aesthetic appreciation are very different properties, and as such it seems emotions cannot be appropriated as an indirect measure of beauty.

In Western culture major chords and scales are considered ‘happy’ whilst minor chords and scales are considered ‘sad’. It is clear that this is cultural, since these particular chords and scales do not appear universally. In medieval music the scale is different and to a modern ear sounds perhaps bittersweet, but it would have had no strong valence to those who first heard it. Little music research actually covers non-Western (or non-modern) cultures so making comparisons is difficult.

One common way to represent this is the valence/arousal diagram seen in Figure 3.1.

Arousal refers to ‘excitement’ and valence, like in electronics or chemistry, means ‘positive or negative’. The general perception is that fast music is high arousal, slow is low arousal, major key is happy, minor key is sad. This only works so well within Western culture, which is of course influential across the world. The arousal scale is perhaps more transferrable than the valence

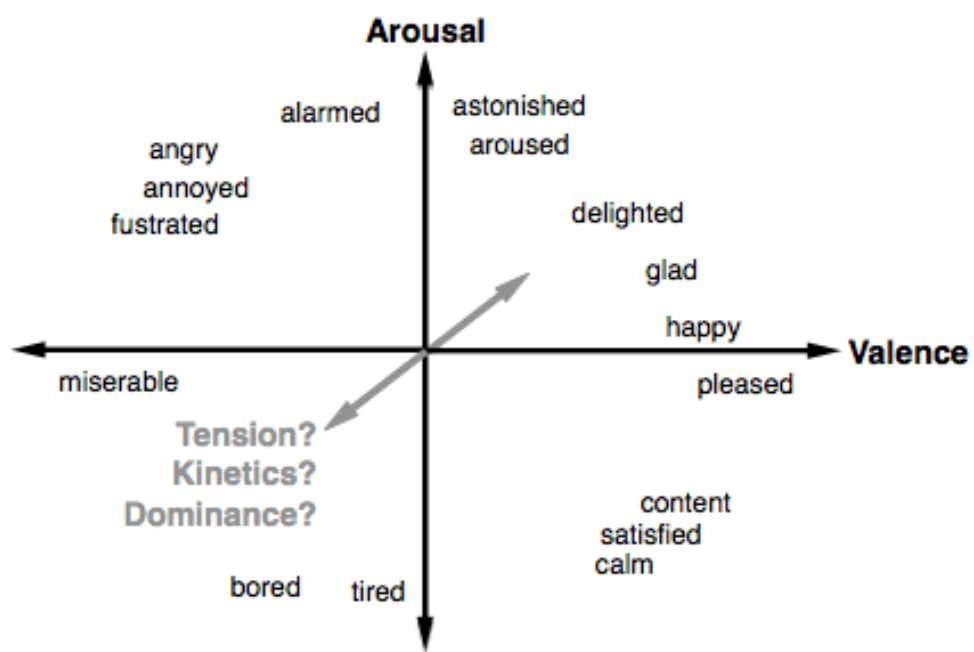


Figure 3.1: Valence-Arousal diagram from Youngmoo paper [Kim et al., 2010].

scale in terms of applying regardless of culture.

Chords and Rhythms

Burgoyne *et al.* compared various ways of automatically recognising chords, but found that none were completely satisfactory [Burgoyne et al., 2007]. The authors favoured Pitch Class Profiles (PCPs) which are vectors based on a 12-chroma scale - all values are non-negative and sum to 1. Dirichlets are a common multinomial distribution which behave similarly to PCPs and with similar constraints, so these were also tried. Hidden Markov Models were considered. Conditional Random Fields performed well, but training takes more time. CRFs with Dirichlet distributions performed impressively. Work is also being done on rhythm recognition [Païement et al., 2008]. Tempo recognition is mostly solved, but true rhythm recognition, with strong and weak beats and long-term dependencies, cannot be solved by current statistical methods alone. Païement presents a novel generative model which represents the structure as dyads (pairs), dyads of dyads, and so on. It appears to function even when given music with non-powers of 2 as the number of beats, because at higher and lower levels dyads are still prevalent. It out-performs Hidden Markov Models (HMMs) on the same task.

One obvious problem with chord recognition is that it implicitly presumes a Western-type scale and associated rules. A chord of C major means nothing in the contexts of, for example, Inuit throat singing, or Australian aborigi-

nal music. Much “world music” has very different rules and conventions to Western music. The same can be applied to some descriptions of rhythm; indeed, it has been shown that Western listeners can interpret unfamiliar music differently in rhythm to the performer’s intention or interpretation [Cross, 2001]. The work of Fujunaga and others is therefore strongly grounded in certain types of music and cannot be easily applied elsewhere.

Establishing Ground Truth

Up until recently most ground truth for machine learning of musical aspects was done by a survey of volunteers or paid participants who would listen to music and provide whatever information was being investigated, e.g. genre, or mood, or which pieces are most similar. Recently the growth of online communities with a wealth of user-provided metadata has led many investigators to take information already existing rather than spending time eliciting responses from participants.

One paper in the Journal of New Music Research’s special issue “From Genres to Tags: Music Information Retrieval in the Age of Social Tagging” looked at using machine learning to generate social tags from MP3 files [Bertin-Mahieux et al., 2008]. The tags were obtained from social network Last.fm. Boosted weak classifiers were used for the categorisation. Boosting is an example of an ensemble method – a process by which several classification algorithms can be combined to create a system which performs better than

any individual classifier. It is usual for ensembles such as this to combine weak classifiers, that is, classifiers which perform a little better than random guessing. Tags were considered more versatile than any single other descriptor, since tags can be of genre, mood, opinion, style, or context-based – such as the film in which the song was featured, or the person the song was written about. Badly tagged songs could be automatically fixed by such a system if successful. It would also provide a means to bootstrap any new site which in the beginning would have very few tags and the tagging system would be virtually useless until a critical mass of users is reached. A system such as this could put in a certain amount of data to start with and leave humans to add and correct tags as necessary.

Features were taken over 100ms excerpts and then aggregated as distributions such that there were 12 features per song, covering a 1 minute extract. The weighting attributed to each tag was relative to other tags applied to the same song, rather than absolute values, because of the disparity in the number of tags per song. AdaBoost [Freund and Schapire, 1997] was used on decision stump weak classifiers. Decision stumps are very simple decision trees, with only one test determining the result rather than several. AdaBoost improves the performance by combining these weak classifiers together. The final tag evaluation, e.g. ‘rockness’, of a song was determined as the number of ratings in the high range, minus the number of ratings in the low range. The middle range was therefore ignored. The performance was judged against listener correlation on the same songs. The result was not as good as real social

tagging, but using both real and computer tags give better predictions than either alone. One problem with this is the popular Western music bias of the source data.

However, it is a current and past problem in MIR that non-Western MIR research is under-represented [Proutskova, 2007]. Proutskova hypothesises this is because non-Western music is hard for MIR people to find, even though archives do exist. These archives are in varied formats and may be in bad condition or of poor quality. Helpfully, social context is usually preserved alongside the recordings. Unfortunately digitization of these archives ranges from 1% - 50%, and not all is freely available online to researchers. There are additional difficulties relating to search criteria assumptions. Western music is categorised by composer, artist and album, whereas non-Western music recordings may have never appeared on an album, and the composer, artist, or both, may be unknown. This makes non-Western music difficult to search for. More usual criteria for this sort of music is a search by culture, place, language, or purpose (e.g. dance accompaniment, or sacred music). Some collectors and some countries or states restrict access to archives they hold.

Another new source of data is the web-based game. A game called MajorMiner was designed by Mande and Ellis, for obtaining objective and specific descriptions of songs. Participants listen to 10 second excerpts and get points for descriptions that match those of other participants, unless they

have been made so many times that the observation is not useful. This encourages some originality. Points were awarded when other players agree with what a participant has said, as well as the reverse. The results were found to be better at classification by a modest amount than were pre-existing tags [Mandel and Ellis, 2008].

3.3.4 Geospatial Analysis

This section addresses techniques in representing topological data from the surface of the Earth. Spatial statistics are relevant to Chapter 5 where musical track's locations of origin are predicted.

The term spatial statistics refers to the application of statistical concepts and methods to data that has an explicit spatial structure which is important to understanding those data [Ripley, 2005]. Typically, spatial data samples are not independent, unusually for traditional statistical methods [Beguin and Thisse, 1979]. Uncertainty in spatial data is common, either from limited instrument capabilities, vagueness of terms, or missing data.

The most common applications are in statistical geography but its use in epidemiology is also well-known. One common example in statistics textbooks is the early work of John Snow who in 1855 proved that cholera was waterborne via statistical means and a geographical dot map showing the occurrences of cholera were clustered around a particular water pump. On

a larger scale, geographic information systems including global positioning systems in recent years have created the discipline geospatial information studies wherein large databases of geographic information are analysed using geospatial relationships such as adjacency, containment and distance.

With the growth of the internet more and more spatial resources have become available to the wider world at low or zero cost. Recently and in particular, the Google Maps API (Application Programmer Interface) created the ability for anyone with access to annotate maps of anywhere in the world at will. NASA (National Aeronautics and Space Administration), an agency of the US government, has been a useful resource for open geographical data for much longer. The founding legislation of NASA written in 1958 includes the phrase ‘provide for the widest practicable and appropriate dissemination of information’. One dataset made available by NASA is the land mask, which gives each square of a gridded representation, at various resolutions and thresholds, a value representing whether that square is land, sea, or coast. This is used in Chapter 5.

Traditional statistical techniques are now being complemented by the application of machine learning techniques to spatial data, but often the spatial information is not used [Gahegan, 2000]. Instead datapoints are sometimes treated like any other numeric values, ignoring the distance information inherent. Gahegan emphasises that spatial information should be used where possible. Neighbourhood classifiers can offer considerable benefits when ap-

plied to data with such structure [Jeon and Landgrebe, 1992]. The way that the space is represented is important – representing each dimension separately makes it more difficult for the algorithm to learn the underlying structure. kNN, Neural Networks, Support Vector Machines, and Self-Organising Maps have all been applied to geospatial data [Kanevski et al., 2008]. Kanevski *et al.* used the Minkowski distance for their learning algorithms, which is a generalisation of both Euclidian distance and Manhattan distance [Kruskal, 1964]. Euclidian distance is what is measured with a ruler on a flat surface, whereas Manhattan is that with you get if you may only move in one of two orthogonal directions to reach a destination.

Work in Self-Organising Maps (Kohonen nets) refers to the formation of geospatial shapes to avoid the edge problem: the way in which a flat map of a sphere makes the extreme left appear to be very far away from the extreme right, whereas in reality they are very close in the sphere itself. It is especially important, therefore, that any representation is able to “wrap around” such that the largest possible distance is only halfway around the world. One example program which accommodates for spherical shape is GeoSOM [Wu and Takatsuka, 2005]. Since longitude’s zero meridian is arbitrary it is important to remove the influence of its position in this way.

Chapter 4

Beauty Experiments

4.1 Introduction

This chapter addresses the measurability of musical beauty and the extent to which other aspects of music can be predicted with various machine learning techniques. The bulk of the chapter is a bank of tests of basic learning algorithms on various music classification tasks, including a benchmark test used by experts in the field of Music Information Retrieval. Those techniques which fared well in genre tests—a vague and human-defined measure—may be more useful for the automatic prediction of musical beauty than those which did not.

4.2 Facebook Survey

A survey to find out what people think about musical beauty, with specific musical examples to rate, was designed. The main point was to determine if people can agree about musical beauty at all—if this is not possible no further work can be done to predict beauty as humans perceive it. This was implemented on the Facebook platform (<http://www.facebook.com>) because the social nature of Facebook might cause the survey to spread from friend to friend so that more participants were found. To this end some advertising was purchased to encourage Facebook users to help. Tracks were paired for comparison, and users achieved different virtual trophies depending on how many pairs they had rated, hopefully encouraging some competition and a feeling of achievement. Later a web survey with identical content was created to allow those who were not members of Facebook to participate. The surveys take some demographic information from the volunteer and then proceed to ask them about paired extracts of music. They are asked to say which extract is more beautiful. 146 extracts of various pieces by Bach, Beethoven, Brahms, Handel and Mozart were hand-cut from MP3 originals (the sound quality was not good as some of these were from archived wax cylinder recordings, so it is unlikely any further problems were introduced by MP3 compression!) The result was 3 extracts per piece taken from the beginning of each piece, each extract being about a minute long but having a “sensible” beginning and end, as judged by the author. All the music

used is out of copyright because it was not only written over 75 years ago (by some margin) but performed over 50 years ago. This was to ensure no copyright law prevented the pieces being distributed over the internet. It has since been understood that short extracts for the purposes of research are an exception to copyright law. The music was unfamiliar at least to the author, and probably unfamiliar to anyone who is not a classical music lover. Still, some people will no doubt have known some of the music, which is a confounding factor. The Facebook survey pairings for the extracts were randomised in both what pairings were made, and which order they appeared to the user each time, to avoid bias in the user picking e.g. the first example presented by preference. The randomisation of the pairings was done as a simple means to avoid the inherent dependence in the order of the extracts filenames to which piece they came from. 73 independent pairs were created. After any pair reached 40 responses, it was retired, to avoid wasted effort – 40 responses are enough to test for statistical significance. All retired pairs were checked for statistical significance using the one-tail binomial test [Siegel and Castellan, 1988], including the Bonferroni correction [Miller, 1966] since there are several simultaneous tests. Bonferroni adjustment accounts for the chance that with enough pairs rated eventually one of the pairs would appear significant owing to the random distribution of raters. As can be seen in Figure 4.1, there was no option to say that the beauty was equal. This was to discourage lazy behaviour on the part of participants - they should try to find a difference in beauty between the two tracks if possible. The

number of participants choosing randomly was expected to be too small to achieve significance in combination with choice ordering, unlike the effect from music ordering which applies to all raters of a pair. In any case, a random significant result would only introduce noise, making classification harder, rather than overestimating performance.

facebook

191

Search

Compare Music!

You will need sound for this! Please choose from the two extracts presented which one you think is **more beautiful** than the other. Click the appropriate radio button and then click submit. If you cannot decide, please **pick one at random** -- don't worry, this will be accounted for in the statistics. Your place will be saved for next time if you leave. All data is saved as soon as you press submit.

▶

Song: [First](#)
Artist: [First](#)

▶


Song: [Second](#)
Artist: [Second](#)

☐ First ☐ Second

Submit

You have rated 56 pairs.

You have attained the rank of **Trombonist**. Rate 4 more to attain the next rank.



[Amend your personal information](#)

[View your results and invite friends](#)

Figure 4.1: Facebook App for comparing musical beauty

4.3 Results

Before running the experiment the proportion of statistically significant results was estimated using some simple assumptions: 1, that the distribution of beauty in music fits a normal distribution, and 2, that people’s ability to rate music for beauty was also distributed normally. Combining these distributions with values for the number of pairs and the anticipated number of raters leads to the estimate of 25% statistically significant pairs, that is, those pairs which are both different enough in beauty and were rated by people who were good enough to distinguish that. The assumptions are extremely broad in the estimate and it was done to give a rough idea of how many raters and songs should be used to get enough rated pairs. With 146 extracts – 73 pairs – 25% would give us 36 useable tracks for analysis. Unfortunately, the results of this survey came much more slowly than had been anticipated. Most participants rated only 10 pairs, and only 12 pairs had result sets full enough for statistical analysis after 10 months uptime. Paid volunteers were considered for a future survey. Something else learnt from the Facebook application was the attention span of listeners – many complained that 1 minute was a very long time to have to listen, and claimed that their decision was made before the end of the excerpt. With this in mind it may be a good idea to represent this fall in attention in the machine learning, such that information from the start of the track is treated as more important than subsequent information, or perhaps discarding most of the track information.

ID	A wins	B wins	N	1-tail test
5	32	38	70	0.2752
10	31	12	43	0.0027
19	20	47	67	0.0007
32	16	37	53	0.0027
49	20	26	46	0.2307
63	27	14	41	0.0298
64	13	31	44	0.0048
65	17	26	43	0.111
66	16	22	38	0.3679
67	15	17	32	0.43
68	30	9	39	0.0005
69	22	13	35	0.0877
70	11	14	25	0.345
71	14	8	22	0.1431
72	7	14	21	0.0946
73	1	18	19	0.00004

Table 4.1: Pairs rated and significance tests

The “A wins” column contains the number of votes for one track of each pair, and the “B wins” column contains the number of votes for the other track. This is, of course, the same track A and the same track B each time, though the tracks were randomised for the participant. Values in the 1-tail binomial column below 0.00625 indicate a statistically significant result at the 10% significance level with the bonferroni adjustment applied. Roughly a third are significant, which is a greater proportion than predicted in the original analysis, but is still not enough examples to train a machine learning algorithm to any credible level for representing beauty. Emboldened in the beauty columns indicates the perceived winner of such examples. One result of particular note is the pair with ID 73. 18 votes to 1 go against

Haydn’s Toy Symphony (ID 73), which is something of a cacophony as it contains many parts played on children’s toys including rattles etc. It makes sense, therefore, that the other example – a piano concerto by Mozart – was defined more beautiful by an overwhelming majority, perhaps simply for its lack of dissonant and sometimes grating noises. Another section of the Toy symphony was found in pair ID 64, which also lost convincingly against a different Mozart excerpt.

4.4 Learning Algorithm Experiments

4.4.1 Experiment Environment

The environment for conducting experiments incorporates code from others in the field with my own scripts for automation. The experiments were conducted through Tacet, MARSYAS, and Weka. Tacet is a scripted program the author wrote previous to the PhD, but it has undergone minor changes since to cope with the difference in data and newer versions of MARSYAS. It automates the overall experiment process from beginning to end. First any MP3, FLAC etc. compressed files are converted into PCM wave files suitable for audio feature extraction. Such extraction is then performed by the MARSYAS program bextract, which has several options for features. For these experiments MP3s were used, in contrast with the subsequent geographic dataset. This is justified by the successful application of machine

learning to genre recognition in MP3s as mentioned in Chapter 3, and the use of MP3s for the benchmark dataset.

Following the feature extraction, which produces an ARFF file suitable for Weka analysis, the chosen machine learning algorithms are run, usually with some cross-validation to achieve training and test sets that are robust to variation. The output of these can then be analysed either within Weka or separately.

4.4.2 Richard Thompson Detector

The next experiment was on a larger dataset but a choice of only 2 classes: ‘Richard Thompson’ or ‘not Richard Thompson’. Richard Thompson is a prolific songwriter, and most of these examples have him on the vocals, but not all. There were 611 examples of which 99 were Richard Thomson tracks were presented to many different machine learning algorithms. The purpose of this experiment was to determine which algorithm would likely perform well with the final data. The default settings for each algorithm as found in Weka were used. As the top algorithms performed so well and so similarly that it was statistically impossible to distinguish between them, a harder problem was set where there were 99 Richard Thompson tracks and 99 other tracks. This experiment still failed to show the difference at the level that was required, so this dataset was no longer used.

Algorithm	n=611	n=198
Bagging REP Tree	91.16%	79.59%
Decorate J48	91.16%	81.63%
Functional Tree	91.16%	81.12%
Multilayer Perceptron	92.31%	81.12%
Rotation Forest	91.33%	82.65%
Simple Logistic	91.82%	81.63%
SMO SVM	91.98%	82.14%

Table 4.2: Results of Richard Thompson Detector - best performing algorithms

Table 4.4.2 shows the results of these trials. Bagging [Breiman and Breiman, 1996] is an ensemble classifier in this case applied to reduced-error-pruning (REP) decision tree weak classifiers. In this instance no pruning was done. Decorate is a meta-learner for building diverse ensembles of classifiers using specially constructed training examples [Melville and Mooney, 2004]. J48 is a type of decision tree which can handle continuous and missing data, based upon the C4.5 tree. Functional trees are classification functions which can have logistic regression functions at the nodes [Gama, 2004]. The Multilayer Perceptron is a feed-forward ANN as described in Chapter 2. The Rotation Forest is an algorithm that creates a diverse and accurate ensemble of decision trees [Rodriguez et al., 2006]. Simple Logistic builds linear logistic regression models [Landwehr et al., 2005]. The Sequential Minimal Optimisation (SMO) SVM is an efficient SVM implementation, here the polynomial kernel was used. Other parameter options were not tried, but this would be a useful addition to the work.

4.4.3 Benchmarking

Each year in the Music Information Retrieval Evaluation eXchange (MIREX) competition researchers test their algorithms against this benchmark for the MIREX prize. The MIREX exercise is attached to the annual ISMIR conference. The competition has several different categories, one of which is genre recognition. A benchmark training set is provided online for pre-competition comparison. The 2004 dataset was chosen as it has been widely used in MIR publications since the exercise, even though the competition itself has moved on to larger datasets. The benchmark training set of 729 examples from the 2004 MIREX competition is used here to investigate firstly which machine learning algorithms perform well at genre recognition, and secondly which feature representations are most attuned to such a task. The results of these experiments give the learning algorithms and feature representations most useful for the beauty recognition task.

4.4.4 Design

The genre of each of the benchmark tracks was predicted via leave-one-out cross-validation, using the default MARSYAS set of features and trying various machine learning algorithms. The best performing algorithms were retried with a larger set of features from MARSYAS (all possible features). The algorithms were chosen to represent a wide variety of classifica-

tion paradigms.

Algorithm	Type	%correct	test measure	%norm	classical	electronic	jazz	metal punk	pop rock	world	n
AllCorrect		100.00%	196.7229081	100.00%	320	115	26	45	101	122	729
Perceptron	Function	66.67%	145.569273	74.00%	256	103	9	17	46	55	486
Simple Logistic	Function	66.26%	144.6707819	73.54%	257	104	14	17	46	45	483
SMO SVM	Function	63.79%	143.3607682	72.87%	258	108	7	14	38	40	465
END nested Dichotomies J48	Meta	63.37%	141.0576132	71.70%	253	102	10	15	38	44	462
Ensemble Selection	Meta	59.53%	138.9396433	70.63%	256	101	26	14	30	28	455
Logit Boost Decision Stump	Meta	61.45%	137.7969822	70.05%	248	103	26	21	26	41	465
Decorate J48	Meta	60.77%	136.175583	69.22%	246	103	9	14	39	32	443
Bagging REP Tree	Meta	59.95%	135.8052126	69.03%	247	107	8	15	26	34	437
Dagging SMO SVM	Meta	59.81%	135.7160494	68.99%	247	112	5	14	39	19	436
Classification Via Regression M5P	Meta	59.67%	134.3058985	68.27%	244	102	10	16	27	36	435
Random Forest	Tree	59.26%	133.7613169	67.99%	242	107	7	14	29	33	432
RBF Network	Function	59.26%	131.739369	66.97%	237	100	10	15	37	33	432
IB1	Lazy	61.59%	130.9245542	66.55%	225	106	11	17	37	53	449
IB k	Lazy	61.59%	130.9245542	66.55%	225	106	11	17	37	53	449
Multi Class Classifier	Meta	58.85%	130.4636488	66.32%	232	95	6	13	44	39	429
Kstar	Lazy	60.36%	129.2921811	65.72%	223	109	9	15	38	46	440
Bayes Net	Bayes	60.71%	128.7942387	65.47%	225	92	17	22	35	52	443
FT	Tree	59.26%	128.5130316	65.33%	228	89	12	19	44	40	432
Naive Bayes	Bayes	56.24%	121.1865569	61.60%	215	94	16	19	28	38	410
LWL	Lazy	48.01%	102.3072702	52.01%	177	97	0	17	56	3	350

Table 4.3: Benchmark Data Testing - learning algorithm comparison, standard featureset, best performing

The feature representation here is the default set from MARSYAS: MFCCs, zero-crossings, and spectral centroid, flux, and rolloff. All percentages are the percentage of correct classifications achieved from leave-one-out cross-validation testing. The genres ‘metalpunk’ and ‘poprock’ are combinations of other genres: respectively, metal and punk music, and pop and rock music. The first row shows the values for a perfect score. Those in the genre columns are the number of correctly classified entities. These results, presented in Table 4.3 have been normalised in line with the real measure used in the MIREX competition in 2004. The normalisation accounts for the difference in class representation: for example, there are 320 examples of classical music but merely 26 of jazz/blues. The top value, 74%, is not far off the 79% achieved by the winning team in 2004, and here merely default representation choices, and default settings on the machine learning algorithms within Weka, are used. It is notable that logical functions (here SMO is a Support Vector Machine (SVM) with Sequential Minimal Optimisation) and meta-classifiers (those that combine the results of several classifiers in some way) have performed well on the task, whereas e.g. lazy classifiers (which store all of the training samples and do not build a classifier until a new sample needs to be classified) have performed less well. It is possible that using information only from nearby instances in featurespace is too simple to represent these classes. This informs the future algorithm focus of the work.

Algorithm	All features	Timbral features	test measure	ALL features	Timbral features	class	elec	jazz	metal punk	pop rock	world	total
ALL COR-RECT	100.00%	100.00%	196.7229081	100.00%	100.00%	320	115	26	45	101	122	729
SMO	68.18%	63.79%	147.648834	75.05%	72.87%	258	104	10	18	48	59	497
Simple Logistic	67.49%	66.26%	145.5308642	73.98%	73.54%	254	102	9	21	49	57	492
Logit Boost Decision Stump	63.24%	61.45%	143.2578875	72.82%	70.05%	260	106	5	17	36	37	461
END nested Dichotomies J48	63.10%	63.37%	141.8916324	72.13%	71.70%	256	100	6	16	41	41	460
Dagging SMO	62.41%	59.81%	139.7297668	71.03%	68.99%	251	113	4	17	41	29	455
Bagging REP Tree	62.55%	59.95%	139.1796982	70.75%	69.03%	249	111	7	17	34	38	456
Decorate J48	63.37%	60.77%	139.0713306	70.69%	69.22%	245	103	8	14	44	48	462
IB1	63.24%	61.59%	135.1124829	68.68%	66.55%	233	111	10	18	36	53	461
IB k	63.24%	61.59%	135.1124829	68.68%	66.55%	233	111	10	18	36	53	461
Bayes Net	56.38%	60.71%	119.739369	60.87%	65.47%	210	84	17	20	32	48	411
LWL	51.03%	48.01%	113.7901235	57.84%	52.01%	197	104	26	45	52	0	424
Naive Bayes	52.81%	56.24%	111.3196159	56.59%	61.60%	192	83	15	14	35	46	385
Kstar	43.90%	60.36%	108.127572	54.96%	65.72%	205	115	0	0	0	0	320

Table 4.4: Benchmark Data Testing extended featureset, best performing algorithms, comparison with standard featureset

The best performing algorithms were tested against the previous results using an extended MARSYAS featureset that combines all the features MARSYAS provides. One notable absence in the algorithms is the Multilayer Perceptron. There were technical difficulties with Weka for this particular algorithm as the computational power needed exceeded the capabilities of the computer hardware. The additional features are chroma: a measure of fitness to a particular chord, Spectral Flatness Measure: a measure of dullness in sound, Spectral Crest factor – a measure of brightness in sound, Line spectral pairs – a way of representing LPCCs differently, and LPCC or linear prediction cepstral coefficients, the linear analogue to the MFCCs which are calculated against the Mel scale of loudness.

The MARSYAS columns are the percentages with all the features used, the Timbral columns are the results from the previous experiment for comparison (the default features MARSYAS extracts being timbral in nature). The normalised results in columns 6 and 7 clearly indicate an improvement for most algorithms when using the extra features. The algorithms which performed worse when given this extra data are likely unable to ignore or de-prioritize some information in favour of other more pertinent information, and this leads to worsened performance.

Indications here are that despite the change in representation it is still the logical functions and meta-classifiers that are best suited to this task, reinforcing the evidence from the previous experiment. None of these yet beat

the result achieved by the winning team in 2004, but with suitable tinkering it should be possible. It is hopefully reasonable to assume that an algorithm performing well on this task would do so in beauty recognition, since both are human-applied labels to audio sounds that have some influencing aspects within, and some outside of, the audio content itself.

4.5 Last.fm

Last.fm [Last.fm, 2013b] is a social music radio site where users can specify a genre or artist and their personal radio station will provide music that matches their selection. This is facilitated mostly through tagging. Users attach categories to music tracks such as “rock”, “smooth”, “female singer” and thousands more. Being entirely user generated there is a fair amount of noise in this set of tags but the most useful tags (genre, mood-related) are happily also the most used tags [Last.fm, 2013a]. One tag used extensively is the tag “beautiful”. The idea behind this experiment is to use this tag to build a small corpus of beautiful music as judged by a crowdsourced audience from last.fm, and then test if it can be recognised by computer.

4.5.1 Method

The 40 tracks which top the ‘beautiful’ tag were downloaded to represent the beauty class. These have been most often tagged beautiful by listeners

compared with all the music on last.fm. 40 tracks selected from those rated highly as ‘awful’, ‘terrible’, ‘horrible’, or ‘shit’ in tagging were downloaded to represent the ‘ugly’ class. This two-class classification problem was then processed by MARSYAS with default features and using the SVM in Weka via leave-one-out cross-validation.

4.5.2 Result and Discussion

With the experiment set up as described, the result was 87% accuracy. This was encouraging as it means beauty can be measured to some extent by computer algorithms. The example data was certainly very polarised and the dataset was small, both factors which make classification easier. It is possible that the algorithm was picking up on noise factors in production over beauty factors, since very noisy music (whether deliberate or accidental) is likely to be rated badly and music rated beautiful by thousands of people seems unlikely to contain too much noise. Further investigation would be required to determine the difference between detecting beautiful music and detecting badly recorded or produced music. This would be a good opportunity for further work.

4.6 Conclusion

Methods for obtaining ground truth from raters and from benchmark datasets were tested and found to be feasible, though surveys may require more incentives for participation. Machine learning algorithms were compared on the benchmark data set and SVMs were amongst the best-performing. Beauty has been predicted on a small selection of Last.fm tags. Two different feature sets from MARSYAS were compared and the larger was found to improve the results when more complex learning algorithms were used. This set of experiments has provided insights into the most useful feature representation, machine learning algorithms, survey techniques and the possibility of detecting beauty by machine. The Last.fm work leads onto the work in chapter 6 and the other experiments provide a basis for representation and machine learning used throughout. Further work not covered in this thesis could include an expansion of the Last.fm experiment to use a much larger dataset, individual consideration of the contribution of each MARSYAS feature, more consideration of parameter settings for the machine learning algorithms, and consideration of class-balancing approaches.

Chapter 5

Geographical Experiments

5.1 Introduction

This chapter considers the factors within music which most distinguish cultures. It is a confounding factor in determining beauty that what is beautiful in one culture may not be to another because of different familiarity with different sounds. Underneath this are the physics, psychology and mathematics of music which remain unchanged no matter the culture. The aspects within the music which cue people to determining the origin of music, are addressed, in particular non-Western music. The final research chapter, 7, is an investigation of the commonalities between geographical discrimination and beauty discrimination. In theory these should be diametrically opposed if beauty can be considered a universal underlying trait. If they are not, then there is

an aspect of beauty that is not independent of culture. Either way something about the nature of musical beauty and ethnomusicology can be learnt.

The work from this chapter was accepted to the ECML-PKDD workshop ‘New Frontiers in Mining Complex Patterns’, held in Bristol in September 2012. This paper was later expanded to a paper in Springer LNAI [Q and King, 2013].

5.2 Geographical Ethnomusicology

The world contains a vast variety of types of music. This music arose as the result of complex geographical, historical, and prehistorical processes. One way to better understand these processes is to analyse the current geographical distribution of music. The study of this distribution is termed Geographical Ethnomusicology. The problem of determining the geographical origin of a piece of music is complicated. Musical forms are rarely pure. Over time they have influenced each other, and many forms of music have travelled far from their point of origin. In particular the influence of western music is nearly ubiquitous. The influence of other forms of music are also widely distributed, for example: Arabic musical influence spread all around the Indian Ocean, across North Africa, to Spain, to central Asia, etc.; more recently reggae has spread from Jamaica, to the UK, Brazil, Mauritius, etc. The question is: given these complications, how well can a computer predict

the geographical origin of a piece of music?

It could be argued that unsupervised spatial clustering methods such as Kohonen nets [Duda and Hart, 1973] would be best suited to such a task. However, the problem with such clustering methods is that there is generally no objective measure of success. Such methods could find groups of similar music in terms of their audio descriptors, but they would not necessarily extract those features most suited to predicting a location. This contrasts with supervised methods, where the labels on known examples (classes or numbers) enable the objective measuring of whether a method is working or not - does it predict well or badly? As the geographical location of origin of the music is known (to some degree) in the corpus, this information should be exploited. The problem is therefore cast as that of training a machine learning program to be able to predict the geographical origin of pieces of music, i.e. the computer learns a functional relationship between the audio content and its geographic origin on the globe. This predictive task is possible to some extent by human musicologists.

5.3 Method

5.3.1 Music Collection

The corpus was built from the personal collection belonging to Professor Ross King, consisting of 1,142 tracks covering 73 countries.¹ The music used is traditional, ethnic or ‘world’ only, as classified by the publishers of the product on which it appears. No Western music was included as it is naturally hard to place since its influence is global – what is sought are aspects of music that most influence location. Thus, being able to specify a location with strong influence to the music is paramount. This will form the target function for the learning algorithm. To determine the geographical location of origin we manually collected the information from the CD sleeve notes, and when this information was inadequate we searched other information sources. There are most certainly other options as demonstrated by Govaerts *et. al.* but these have varying levels of accuracy and indeed their ground truth for the experiment was ‘personal knowledge’ or ‘by looking up the origin’ [Govaerts and Duval, 2009]. We did not wish to confound the ability of the predictor with incorrect location information. The location data is limited in precision to the country of origin - we did not have time to try to find out more about each track. In many cases the level of detailed precision possible for

¹The music used is subject to copyright, but the processed data is not. Code and data is available at <http://bitbucket.org/Eskoala/python-geolocation/> as a git clone, or email eskoala@gmail.com.

ascertaining musical origin is arguably not much smaller than perhaps the region of a country, except in certain cases where a community has been extremely isolated.

The country of origin was determined by the artist's or artists' main country of residence. Of course, many artists live in different places throughout their lives, but our aim was to determine the major influence. For example, if a Malian writing Mali music lives in retirement in Paris, we consider the music Malian. We recognise that some music is less linked to countries than cultures, but countries at least have true geographical locations that are measurable - by virtue of having defined borders - allowing our machine learning approach to give objective output. We have taken the position of each country's capital city by latitude and longitude as the absolute point of origin in the beginning. The assumption here is that the political capital is also the cultural capital of the country. This assumption also utilises coarse a priori knowledge about population - most, but not all, countries have a highly populous capital city. Using the capital takes into account country-level population distribution in a simple way without resorting to time-consuming investigations into the exact place each artist spent most time. In the population distribution task we altered this point of origin to the centre of population, or population centroid, of each country, which is a fairer measure that takes into account skewed population distribution and capitals with low populations. Countries are linked to artists, not tracks.

It is clear that the country of *production* may have no bearing on the origin, since many world music CDs are made in Germany or the US despite the music coming from, for example, Kenya. The artist in question is usually the composer where known, or else the performer where the music is traditional to their home country. If several artists have made a contribution to the music, all substantial (not merely ‘featuring’) contributions are taken into account when deciding if it should be included. Any track that had ambiguous origin – whether because the artist’s own origin was ambiguous, or many artists from different countries collaborated, or the track is a deliberate fusion of styles – was removed from the dataset. There are no “right answers” in such cases. For example, Bhangra music is a fusion of Indian Punjab culture, UK culture and hip-hop. Therefore to try to determine a single geographical location for a Bhangra track would be nonsensical as it has multiple sites of influence. One could suggest that the geographical midpoint of all the influences is the right answer, but this does not fully encapsulate the data and would result in a position for Bhangra music near Volgograd!

Figure 5.1 shows the distribution of tracks per country, for the 20 best represented countries. It can be seen that some countries are much better represented than others, which will affect the performance: for a given country, the more examples there are in the dataset, the more information about that country’s music is known, which will lead to better predictions.

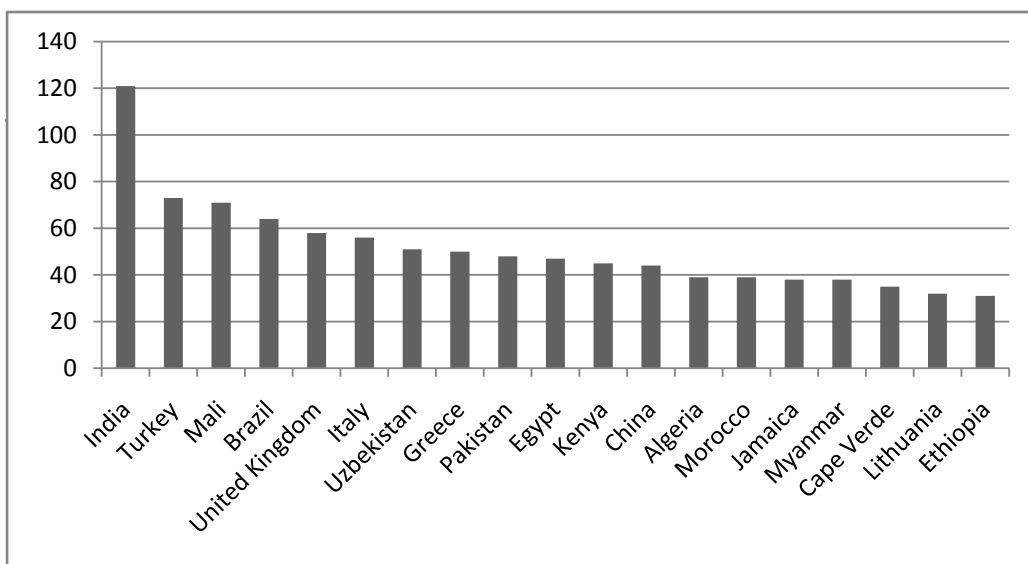


Figure 5.1: Partial sample of music distribution by country

5.3.2 Audio Features

The program MARSYAS [Tzanetakis and Cook, 2000] was used to extract audio descriptors from the wave files. We first used the default MARSYAS settings in single vector format (68 features/attributes) to estimate the performance with basic timbral information covering the entire length of each track. The different approach, using all of the track in this case, is because there is no reason not to use the whole track when human listeners are not involved. The information is there, it might as well be included. In later experiments such as in Chapter 7 only the portion of the track listened to by participants is available to the machine learning algorithm.

MARSYAS Features

The MARSYAS features available are summarised in Table 5.1.

Each of the default features is an indicator of timbre, which is one of the main ways (another being attack-decay-sustain-release models) [Lichte, 1941] to distinguish musical instruments. Since instrumentation is also a major difference between cultural music traditions, these features are appropriate to the task. No feature weighting or pre-filtering was applied. All features were transformed to have a mean of 0 and a standard deviation of 1. We also investigated the utility of adding chromatic attributes. These describe the notes of the scale being used. This is especially important as a distinguishing feature in geographical ethnomusicology – unlike in Western music which largely conforms to one tuning system. The chromatic features provided by MARSYAS are 12 per octave – Western tuning, but it may be possible to tell something from how similar to or different the music is from Western tuning.

5.3.3 Geographic Representation

The problem of predicting a point on the surface of a sphere is made more complicated as the standard coordinate system of latitudes and longitudes, which are the natural targets for regression, have a complicated relationship to surface area - where the predicted point will be. This is illustrated in

the standard Mercator projection of the globe where countries near the poles are unnaturally large. Area is not preserved equally on such projections. In general there is no perfect flat projection – a compromise is made one way or another in favour of a particular desired quality, perhaps straight lines of latitude and longitude, perhaps equal area, even such complicated solutions as the butterfly map by Cahill later re-imagined by Waterman. None of these addresses the true mathematical problem fully – the Earth is not flat, and should therefore ideally be treated as close as possible to its true topology. In considering the sparsity of our data, we choose a sphere as an approximate representation for the globe, though with more precision still it is an oblate spheroid with certain peaks and troughs across the surface.

5.3.4 Spherical k -Nearest-Neighbour Prediction Method

We decided to cast the problem as a regression problem (predicting a point) rather than a classification problem (predicting a country) because the large number of countries, and low number of examples per country, would complicate classification. Most regression methods assume either that only one real number is to be predicted, or if multiple real numbers are to be predicted that they are independent. Perhaps the simplest approach to running regression with spherical coordinates is to side-step this difficulty and use a k -Nearest-Neighbour method to predict points [Duda and Hart, 1973]. A Euclidean (in attribute space) kNN algorithm was run using the musical features as axes.

For each track the nearest k neighbours were found, the geodesic mean of their locations was taken, and the result compared to the true origin. To adopt this method to predict geographical location we used spherical geometry and took the average positions on an idealised sphere of Earth radius, using standard geodesic distance calculations. The results can be measured in terms of their great-circle distance from the true location (capital city) of the piece under consideration.

Finding the geodesic midpoint of the k nearest neighbours: with λ as latitude and ϕ as longitude (both in radians), convert to cartesian coordinates on a unit sphere:²

$$x = \cos(\lambda)\cos(\phi) \quad (5.1)$$

$$y = \cos(\lambda)\sin(\phi) \quad (5.2)$$

$$z = \sin(\lambda) \quad (5.3)$$

Take means of the nearest neighbour points per dimension $\bar{x}, \bar{y}, \bar{z}$. Find the longitude $\bar{\phi}$ of the midpoint:

$$\bar{\phi} = \arctan\left(\frac{\bar{y}}{\bar{x}}\right) \quad (5.4)$$

²In practice a four-quadrant inverse tangent function is necessary for Equations 5.4-5.5 to cover all cases. The function we used is known as `atan2` or `arctan2` in most programming languages.

Find the latitude $\bar{\lambda}$ of the midpoint

$$\bar{\lambda} = \arctan\left(\frac{\bar{z}}{\sqrt{\bar{x}^2 + \bar{y}^2}}\right) \quad (5.5)$$

5.3.5 Utilising *a priori* Background Knowledge

We investigated the utility of using *a priori* knowledge to improve the predictions.

Land and Sea

The first piece of knowledge used is that music is produced on land. To utilise this we applied the NASA LandMask projected onto the idealised sphere of the Earth. This gives the terrain type for each square *degree* latitude by longitude. This varies in true size from about 110km wide to exactly 0 at each pole for longitude, whereas the latitude separation at 1 degree remains roughly 69km apart, excepting differences for the oblate spheroidal shape of the Earth. Because the mask is given in latitude-longitude squares, a line-drawing algorithm must be employed to traverse a spherical distance across the Earth, ensuring that the great-circle calculation is performed between each step, so that the correct next square is chosen. The total land contained in the path of error is weighted against the total water coverage.

Different weightings were investigated but the simplest option was chosen: land 1:0 water such that only the land part counts. The reasoning behind this is that though human migration is slow across land, it is comparatively very fast across water. For example, Brazilian music is close musically to Portuguese music because of migration patterns despite the size of the Atlantic Ocean. Different weightings will be tested in future.

Population Centroids and Population Density

The first change is to recenter each country to its population centroid. The centroids were collated from the GPWv3 per administrative area data [SEDAC and CIAT, 2012] scaled up to per country via the method detailed by Greg Hamerly (<http://cs.ecs.baylor.edu/~hamerly/>). The same dataset also provides population density grids at several resolutions. The coarsest available – per square degree – was used since it matches what we have for the land mask and is thus more easily comparable. Population density (as opposed to count) is chosen because it avoids the problem of varying size of square degree on the earth’s surface owing to the diminishing distance between longitudinal degrees as either pole is approached, as explained above.

The algorithm for applying this mask is as follows:

1. The k nearest neighbours are determined based upon musical features.

2. For the group of nearest neighbours, the geodesic midpoint is found.
3. The geodesic distance to the furthest neighbour from that midpoint gives a radius of a geodesic circle which contains all the neighbours.
4. We calculate a new midpoint for this circle which is weighted by the population density in each of the points, at a resolution of one square degree.

The weighted midpoint is found as in Equations 5.1-5.5 but each dimensional mean is calculated from weighted coordinates such that

$$\bar{x} = \sum \frac{x_i \rho}{k} \quad (5.6)$$

$$\bar{y} = \sum \frac{y_i \rho}{k} \quad (5.7)$$

$$\bar{z} = \sum \frac{z_i \rho}{k} \quad (5.8)$$

where ρ is the population density for the square degree at the point (λ, ϕ) and k is the number of neighbours.

5.4 Results

5.4.1 k -Nearest-Neighbour Performance

Using the default MARSYAS features the best predictive performance we achieved was a 2,827km median distance, and a 2,125km median land distance from the true position. This was achieved with $k=10$.

This result (like all above results presented) is significant at p-value < 0.001 (see below). We chose the median as our main statistic, rather than the mean, as it is more robust to outliers. The results for the means were also each time significant at p-value < 0.001 .

Table 5.2 shows the breakdown of results. The meaning of the columns is as follows: k is the number of neighbours used. The feature set is default (68 features) or default with chromatic (116 features). The mapping is whether the land measure is used to find the closest landmass. Spherical signifies a pure distance measure using spherical geometry. Population signifies the same as spherical but weighted by population. The median is the median distance from the true answer for each combination. The ks shown are a selection but all values from 2-10 were tried, the results of which can be seen in Figures 5.2-5.5.

Whilst demonstrably better than random this result could be improved upon. What is shown here is that even with one of the simplest possible algorithms

the information contained in the features is enough to indicate some geographic information. The land proportion is, of course, always smaller than the total distance, but it was argued that this is a fairer measure of error. This indicates that some measure of population distribution would be useful *a priori*.

It is interesting that for land measurements the $k = 2$ method performed best before addition of chromatic features, but the $k = 10$ version performed best with the extra features. One hypothesis is that for very similar music the chromatic features do not aid the similarity measure already gained by timbral features, yet when more dissimilar music is encountered, chromatic features come into their own as a coarser measure of similarity. Since some countries were underrepresented to the point of having fewer than k member tracks, this follows. However, the simple spherical method medians show no such inversion.

5.4.2 kNN with Population Distribution

Figures 5.2-5.3 compare the simple spherical approach against the population method by median distance error. Figures 5.4-5.5 do the same for the means.

Looking at median, which is fairest for a skewed distribution such as this, for lower k values the use of population distribution leads to an improvement

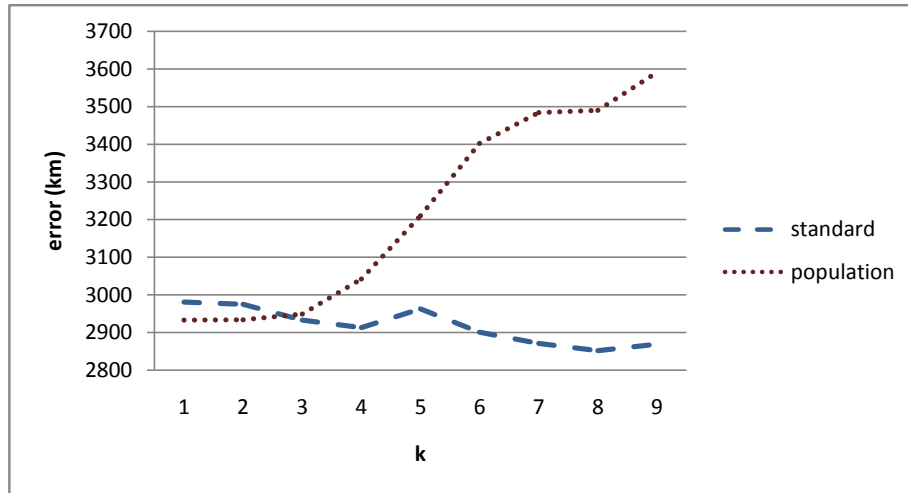


Figure 5.2: Default features performance as median for all k , standard (spherical) measure and population measure.

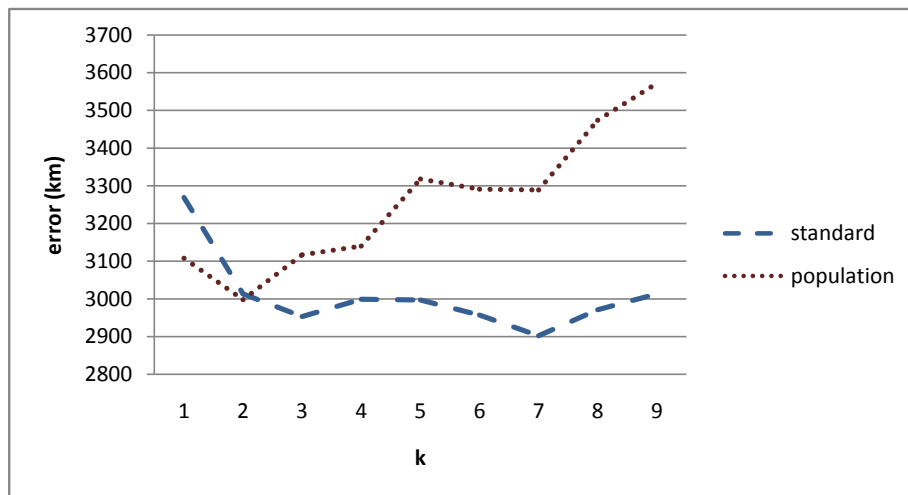


Figure 5.3: Default + chromatic features performance as median for all k , standard (spherical) measure and population measure.

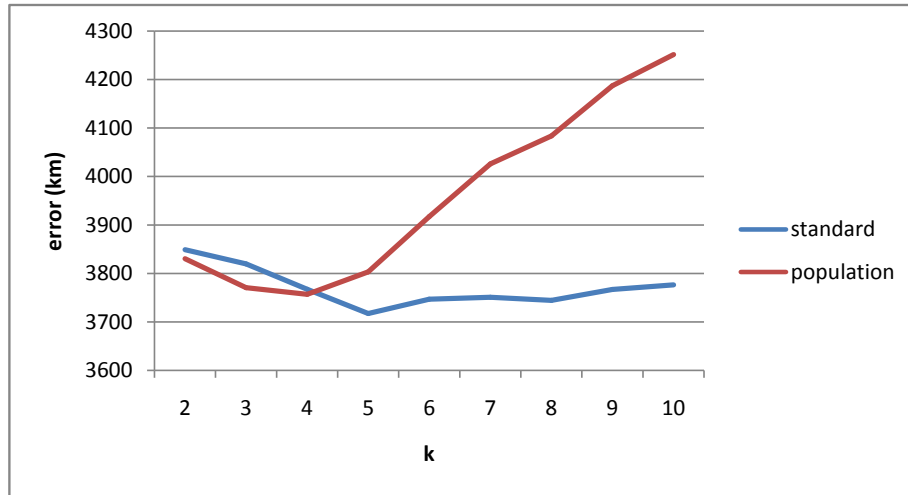


Figure 5.4: Default features performance as mean for all k , standard (spherical) measure and population measure.

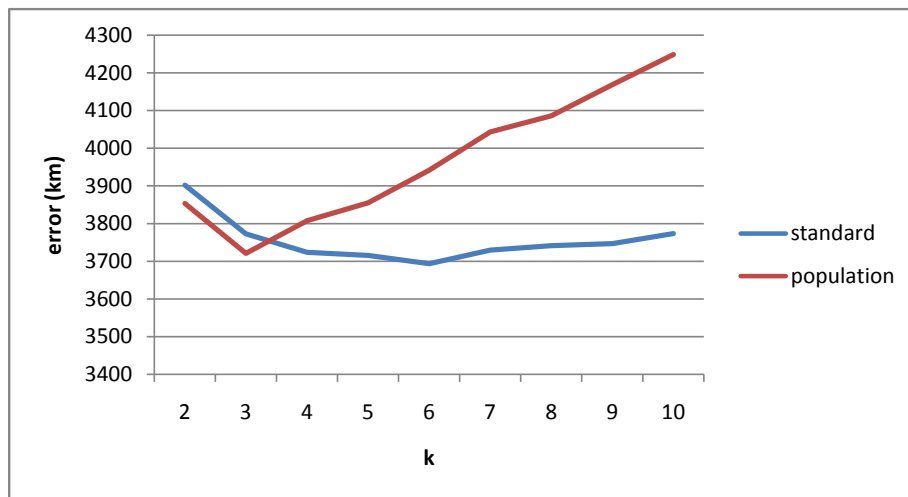


Figure 5.5: Default + chromatic features performance as mean for all k , standard (spherical) measure and population measure.

to results. However, once k is greater than 3 the population method median degrades, and once it is greater than 4 the mean is similarly worsened. Experiments with higher k have noisier data since the additional neighbours are more likely to be from a different country. When more dissimilar music is encountered it is likely that instead of improving a result within a country the method selects more densely populated countries instead.

5.4.3 Statistical Significance

The determination of whether a music geographical prediction method is performing better than random is difficult using traditional statistical methods, because of the complicated geographical distribution of the music and countries. Therefore, we used a computationally-intensive statistical method based on resampling, programming 1,000 random trials for each k and method combination, and measuring our distance means against the distribution of random means to ensure the statistical significance of our results. The random distribution results ranged from 5000km–5500km mean distance error. Our means were all so far left of the distribution that they were significant at $p < 0.001$.

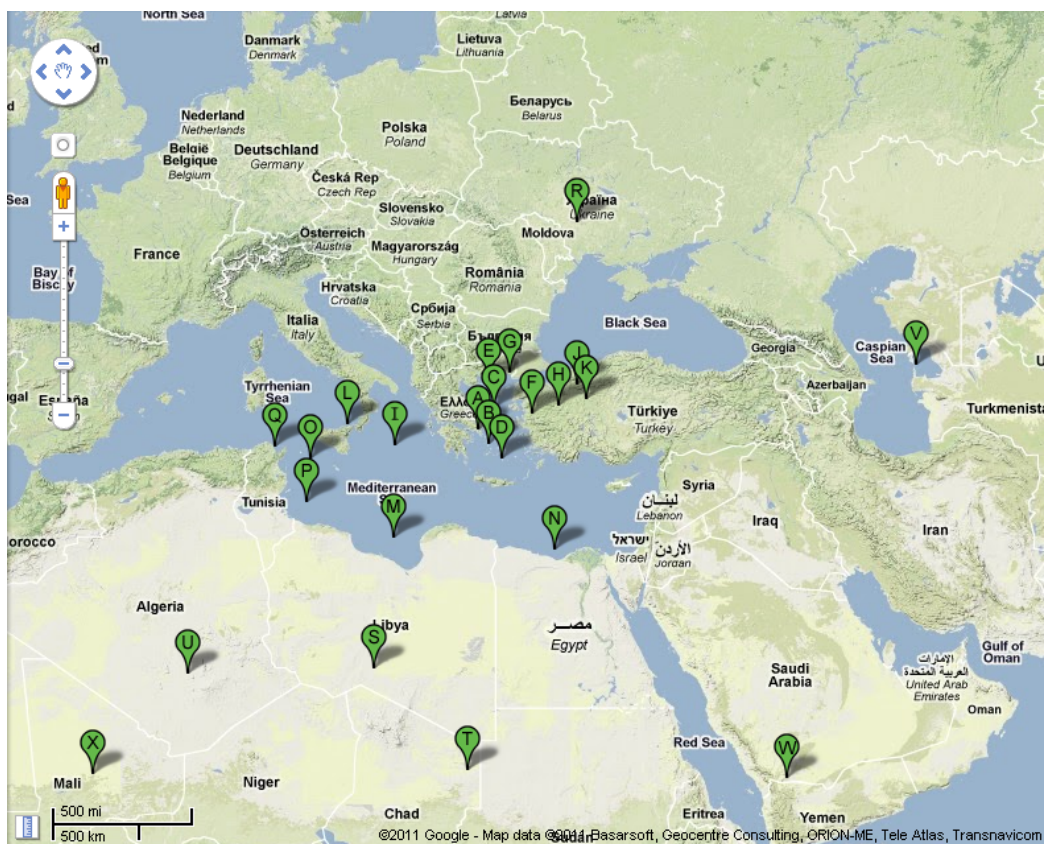


Figure 5.6: Greek music – predictions of location.

5.4.4 Performance by Country

The algorithm performed much better on some countries than others – even with the same number of tracks available, suggesting that some countries are more musically diverse than others. An additional problem is the relative size of countries affecting the level of precision required, for example, Russian music is much more geographically diverse than that of Croatia.

Figure 5.6 (courtesy of Google Maps) shows half of the estimates (for clarity

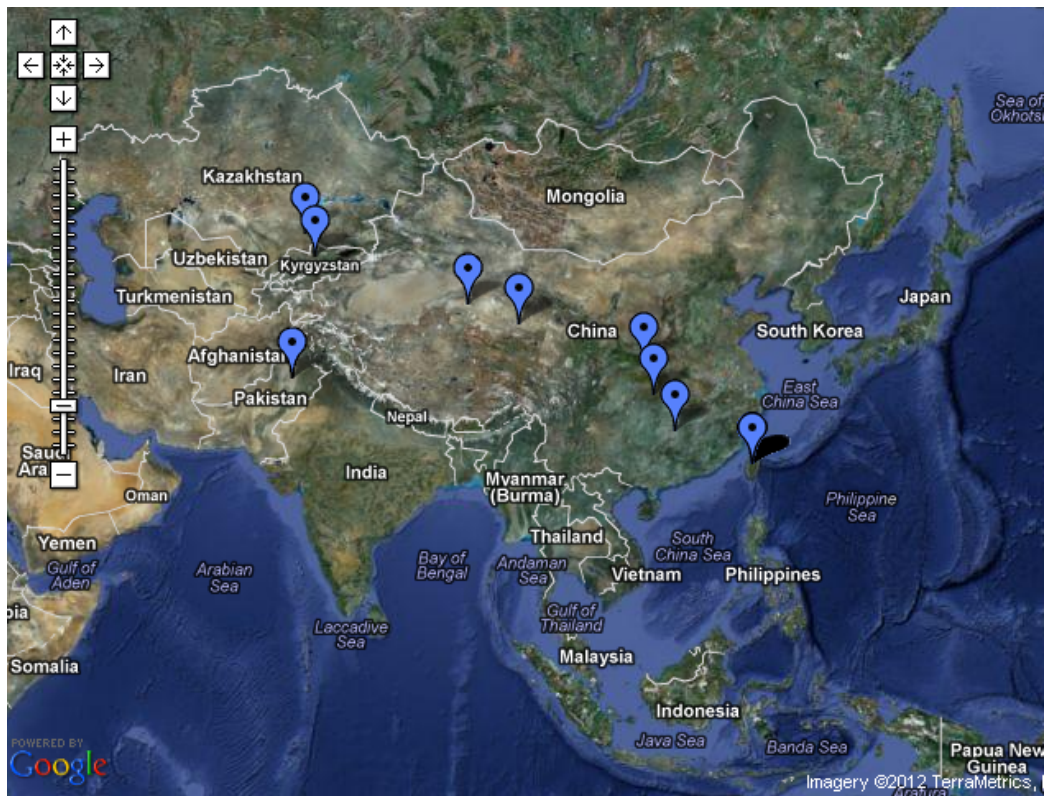


Figure 5.7: Taiwanese music – predictions of location.

and to avoid overloading of positions) for Greek music, taken evenly from the distribution for Greece. From this it is clear that the prediction range for a country can be quite tightly distributed – the furthest estimate is 3,652km but there is a close cluster central to the image that reflects the more general skew in the distribution of estimates for all countries. Figure 5.7 (also courtesy of Google Maps) shows all Taiwanese music, though the correct position is overloaded with 16 predictions. Taiwanese music was found to be particularly recognisable. it is not certain whether this is down to a uniqueness in Taiwanese music or simply a lack of diversity in our Taiwanese examples, however.

5.5 Discussion and Future Work

We have introduced the problem of predicting the geographical origin of music. There is great scope for further research and improvement in prediction performance.

With a larger corpus, with both more tracks from each country and more countries represented, the prediction results will inevitably improve. If one is only interested in predicting location (as opposed to understanding the historical/pre-historical reasons for musical distribution) the problem is in many ways similar to statistical analysis of text, where organisations such as Google have now indexed so many pieces of text that they can solve

many problems that were once thought to require solving deep problems in computational linguistics.

More geographical information could be utilised. It would be better to have access to the exact location of the origin of the music, rather than just the capital or population centroid, as most countries have strong regional variations in style. Some cultures change drastically over small areas, some are unchanged over large expanses, and this needs to be learnt by the prediction method.

Better representations: the music could be better represented for computational analysis. It is a truism within machine learning that the hard part is getting the features correct, and with the correct features almost any learning algorithm will work. For example, extra features such as the fine chromatic feature used by Gomez *et al.* could be applied [Gomez and Herrera, 2008]. It would also be useful to explore the possibility of filtering and pre-selection of descriptors.

It would be interesting to compare the work with a clustering approach to the problem. Since we only have the location to the level of a particular country, classification could be used, though classification with this many classes might prove problematic.

Many other forms of machine learning could be applied: neural-networks, support vector machines, decision trees, etc. It is also generally possible to improve the performance of individual methods by combining them together

to form consensus predictions [Duda and Hart, 1973].

It is difficult to know how good our prediction results are as there are no previously published related comparisons. It would therefore be very interesting to compare the results of the machine learning programs with that of human performance in predicting musical origin.

The motivation for this work is to better understand the diversity of world music. To do this we have to go beyond just the prediction of location, but to analyse what features of the music are responsible for these predictions. This is now the main focus of research. Once these are known, it would be very interesting to attempt to generate music appropriate to a particular region.

5.6 Summary

Traditional research into the arts has generally been based around the subjective judgment of human critics. We propose an alternative approach based on the use of objective machine learning programs. To illustrate this methodology we investigated the distribution of music from around the world: geographical ethnomusicology. To ensure that the knowledge obtained about geographical ethnomusicology is objective and operational we cast the problem as a machine learning one: predicting the geographical origin of pieces of music. 1,142 pieces of music from 73 countries were collected and described

them using 2 sets of standard audio descriptors using MARSYAS. To predict the location of origin of the music a method was developed which was designed to deal with the spherical surface topology, based upon a modified k -Nearest-Neighbour. The utility of *a priori* geographical knowledge in the predictions: a land and sea mask, and a population distribution overlay, was also investigated. The best-performing prediction method achieved a median land distance error of 1,506km, with comparable random trials having mean of medians 3,190km –this is significant at $p < 0.001$.

Feature		Explanation	Grouping
Time	Zero-crossings	The number of times a signal changes sign, <i>i.e.</i> how often it crosses the horizontal zero line.	Default timbral
Spectral	Centroid	A measure of the “centre of mass” of the power spectrum.	Default timbral
Spectral	Rolloff	Describes the amount of skew in the power spectrum.	Default timbral
Spectral	Flux	Is an indicator of the amount of spectral variance based upon differences between adjacent spectral windows.	Default timbral
Mel-Frequency Cepstral	Coefficients	coefficients for a mel-frequency (tailored to human auditory response) power cepstrum - representing the short-term power spectrum.	Default timbral
chroma		detects frequency matches for each musical note of the Western scale (and its octaves)	Chromatic feature
Spectral	Flatness Measure	quantifies how tone-like, as opposed to noise-like, a sound is.	Non-default timbral
Spectral	Crest Factor	Peak to average ratio of amplitude. Variance in loudness across frequency.	Non-default timbral
Line	Spectral Pair	Describe the two resonance frequencies of the vocal tract when open or closed.	Non-default misc
Linear	Prediction Cepstral Coefficients	as MFCC, but linear rather than Mel-scale	Non-default timbral

Table 5.1: MARSYAS features

Table 5.2: Median and mean distance from true location per k , featureset and algorithm

k	Features	Mapping	Median (km)	Mean
10	default	spherical	2869	3776
5	default	spherical	2913	3717
3	default	spherical	2975	3820
2	default	spherical	2980	3849
10	def+chrom	spherical	3013	3774
5	def+chrom	spherical	2999	3715
3	def+chrom	spherical	3012	3772
2	def+chrom	spherical	3269	3902
10	default	population	3591	4251
5	default	population	3042	3803
3	default	population	2987	3825
2	default	population	2933	3830
10	def+chrom	population	3572	4249
5	def+chrom	population	3140	3855
3	def+chrom	population	2997	3721
2	def+chrom	population	3107	3853
10	default	land	2125	2675
5	default	land	2024	2610
3	default	land	1966	2665
2	default	land	1850	2583
10	def+chrom	land	1506	2694
5	def+chrom	land	1550	2613
3	def+chrom	land	2087	2639
2	def+chrom	land	1996	2627

Chapter 6

Beauty in World Music

6.1 Introduction

Having considered non-Western music in Chapter 5, it seemed prudent to use non-Western listeners to further remove bias from the system. With this in mind a survey was prepared to determine how a specific group of non-Westerners rate diverse non-Western music for beauty. This chapter shows a development of the method conducted for the Facebook survey. The ratings go on to be used in a learning experiment to see if the ratings can be predicted by machine.

6.2 Design

6.2.1 Music Used

Music from 93 different countries was presented to participants in a random order, different per participant, in discrete pairs. The extracts (from PCM wave files) are each 20 seconds long and taken from a point 30 seconds into the track. The beginning of music tracks is often quiet and often not representative of the overall feel. 20 seconds is enough time to form an opinion as was reported by participants in the Facebook survey (see section 4.2) and it is important to make the most of participants' time. The tracks were selected from the same set used for the geographic experiments, paired such that they were the furthest distance apart. The aim of this was to compare music of different cultures and see if any underlying influence of beauty in the music could override cultural preferences. The algorithm for selecting tracks is as follows:

1. Create a list of all possible pairs
2. Determine the distance for each pair (cosine rule distance from known location of each track)
3. Order the pairs by distance, greatest distance first
4. Select the first pair, maintaining a list of selected pairs
5. While more pairs are required

- (a) Find the next pair that has no tracks in common with all previously selected pairs
- (b) Select that pair, adding to the list of selected pairs

This is not the optimal solution for this problem, however a full set of non-overlapping pairs was not required, but only a subset of 300 pairs. This method was a good heuristic in terms of both computing time and programmer time. The general solution to this problem may be NP-hard as it combines the knapsack problem [Matthews, 1896] with an extra condition of independence making many combinations mutually exclusive.

6.2.2 Listening Conditions

Twenty PCs with headphones, internet connection and web browser pointed at the survey are provided in a quiet room.

Participants were not told the overall purpose of the study as this might have skewed their responses. The question about happiness is there partly to disguise the specific purpose of the survey and partly to encourage participants to think about the difference between the music being beautiful and 'liking' or 'enjoying' the music. The results are not analysed separately but are useful for some comparisons.

6.2.3 Participation

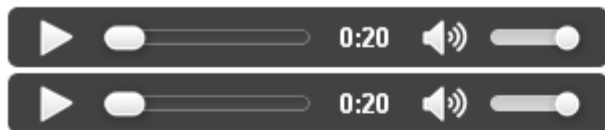
Students in Singapore will be asked to participate in the survey. They were given a login ID and password. The first page asks for some demographic information. An example (fictional) response is shown in Table 6.1. This part was based upon previous marketing studies conducted by our colleagues in Singapore.

Age in years	25
Gender	Male
Race	Chinese
Housing	HDB 3-4 rooms
Nationality	Singaporean
First Language	Mandarin
Second Language	English

Table 6.1: Example demographic information

Following the demographic part of the survey the pair listening test was presented. In this they were asked to rate the tracks against one another in terms of beauty and whether it made them feel happy. Two music players, one per extract, were presented with associated play, pause, and stop controls. There was no mechanism to prevent listening to both tracks simultaneously but this behaviour was not expected as it would be unlikely to help or be pleasant for the listener. See Figure 6.1 for an image of the listening experiment webpage.

Listening Experiment



Answer the following:

If for some reason one or both of the extracts is not working, check this box: ☐

Which extract is more beautiful?

First ☐ Second ☐

Rate each extract on the following scales.

Extract 1

Ugly ¹ ² ³ ⁴ ⁵ ⁶ ⁷ Beautiful
☐ ☐ ☐ ☐ ☐ ☐ ☐

Extract 2

Ugly ¹ ² ³ ⁴ ⁵ ⁶ ⁷ Beautiful
☐ ☐ ☐ ☐ ☐ ☐ ☐

Figure 6.1: Part of the listening page, happy scales are off the bottom of the image.

6.2.4 Grouping of Listeners and Pairs

Each pair is listened to by 20 participants. Each participant is asked to spend 20 minutes on task, resulting in each participant rating about 15 pairs (30 tracks). This is based on 40 seconds listening and 40 seconds answering questions. The overall result will be 150 pairs rated (300 tracks), each rated by 20 participants. This gives about 3 tracks per country. Each participant was part of a group that heard the same 15 pairs but each heard the pairs in a random order. Not only was the order of the 15 pairs randomised, but the order of appearance of the two tracks within each pair to avoid the chance that a preference for always picking the first, or always picking the second, were exhibited. They were each paid \$10 (SGD) for their time once they have completed the task, irrespective of their particular responses.

6.3 Survey Results

The ratings used are the paired ratings for beauty, as agreement is easier to measure for these. All ratings are assessed for the statistical significance of the agreement of raters using a two-tailed binomial test. Agreement was found to $p < 0.05$ for 91 pairs (182 tracks) of the 150 pairs. However, this does not incorporate the Bonferroni correction. When this is applied only 34 pairs (68 tracks) can be considered significant. This does not mean that the information in the remaining pairs has no value to any machine learning

algorithm - though of course the more agreement between raters the more reliable and therefore useful the result will be. The results are taken as 4 possible groups for input in to the machine learning algorithm: significant at 5% with Bonferroni correction (68 tracks), significant at 10% with Bonferroni correction (80 tracks), significant at 5% ignoring Bonferroni considerations (182 tracks), and all rated groups with an overall majority vote (all but one pairs) The significance and the paired percentages for beauty and for happiness can be seen in Appendix A. The scale ratings could be used in a future study. From this it appears that the ratings are tied to the familiarity or closeness of the country they are from. This is unexpected, but it gives scope to test this idea to find the extent of the effect. This forms the basis of Chapter 7. Further work might include looking at these ratings on a per-country basis rather than by distance.

6.3.1 User Demographics

The demographics of the respondents were not as diverse as was hoped. The modal demographic combination of nationality, race, first and second language was Singaporean nationals of Chinese descent with first language English and Second language Mandarin, of whom there were 107 out of a total 189. The full demographic description of each participant can be seen in Appendix B. The gender breakdown was fairly even, being 92 female and 97 male from 189 participants. Figure 6.2 shows the nationality breakdown. It

can be observed that the vast majority of those surveyed were Singaporean, with the next most represented group being Malaysians, of whom there were only 18. In terms of race, people of Chinese origin were most strongly represented (see Figure 6.3). The most common first language was English, the most common second-language Mandarin, both accounting for over half the participants (which could be determined from the modal grouping as well). The first and second language distributions can be seen in Figures 6.4 and 6.5. The participants were all students and age ranged from 19 to 28 in a normal distribution (Figure 6.6). The housing situation was part of a standard form our colleagues in Singapore use, which is a proxy for standard of living. There was a fairly even spread with a mode of 77 for HDB34 which is a 3 to 4 room flat from the Housing and Development Board in Singapore, and very few participants in an HDB12 (a 1 to 2 room flat). It seems that the dataset has no particular bias for this but a comparison with averages across the university, and Singapore in general, would be needed to confirm. Figure 6.7 displays this information.

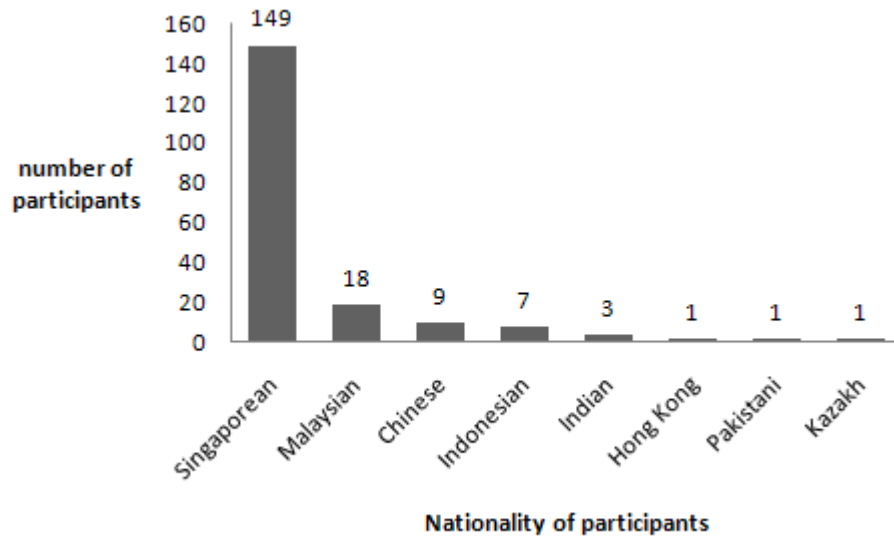


Figure 6.2: Distribution of the self-identified nationalities of the participants.

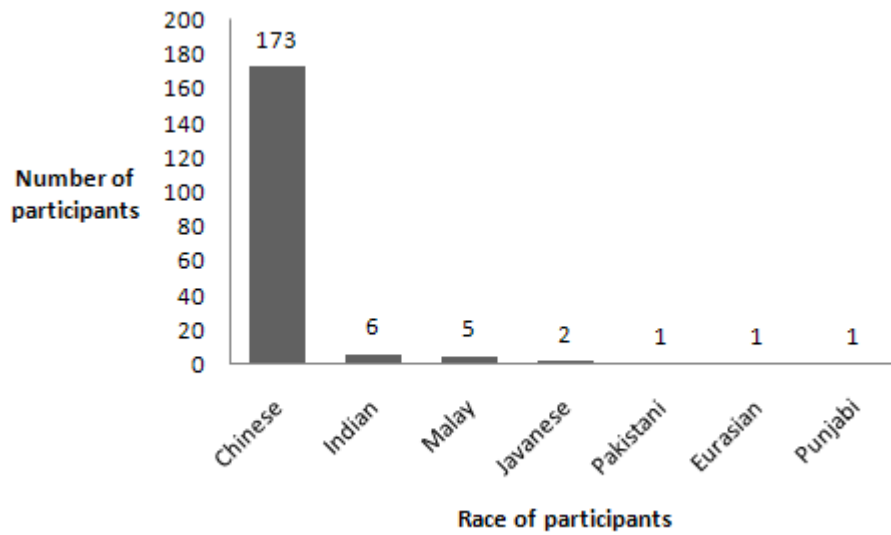


Figure 6.3: Distribution of self-identified race of participants.

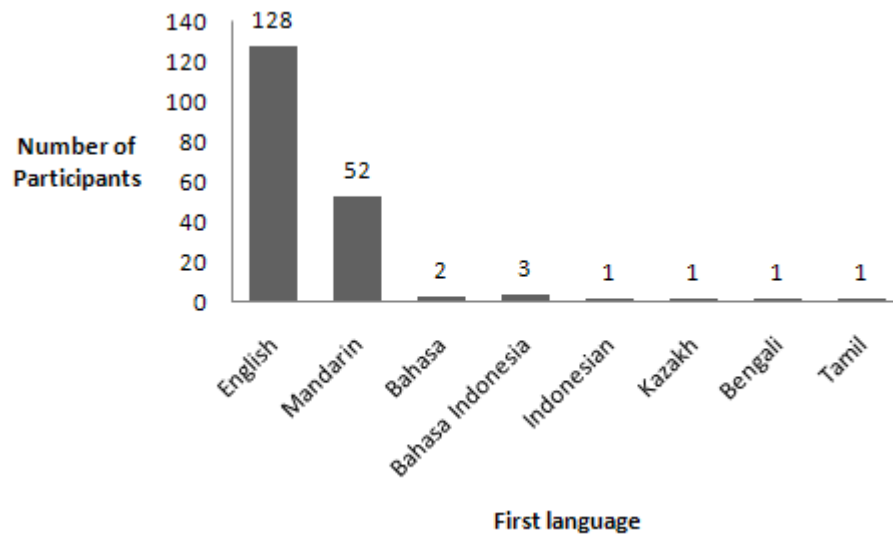


Figure 6.4: Distribution of first language of participants.

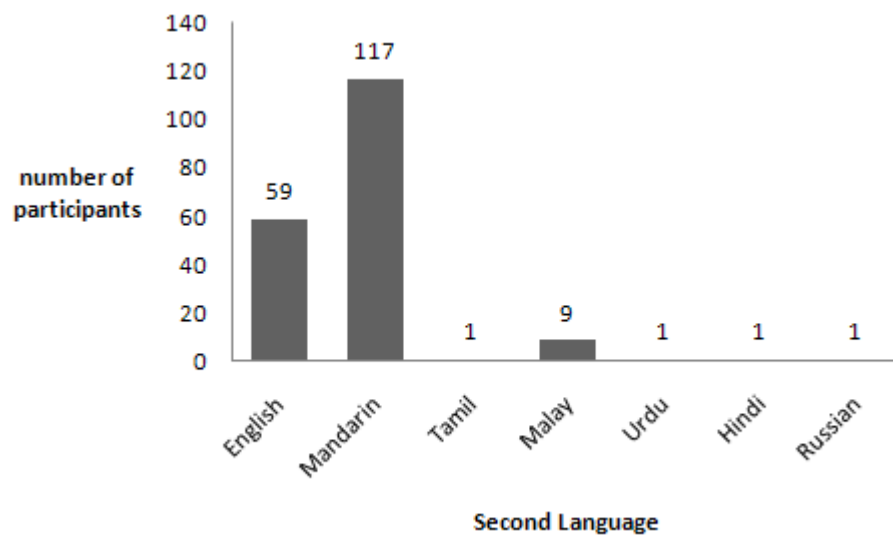


Figure 6.5: Distribution of second language of participants.

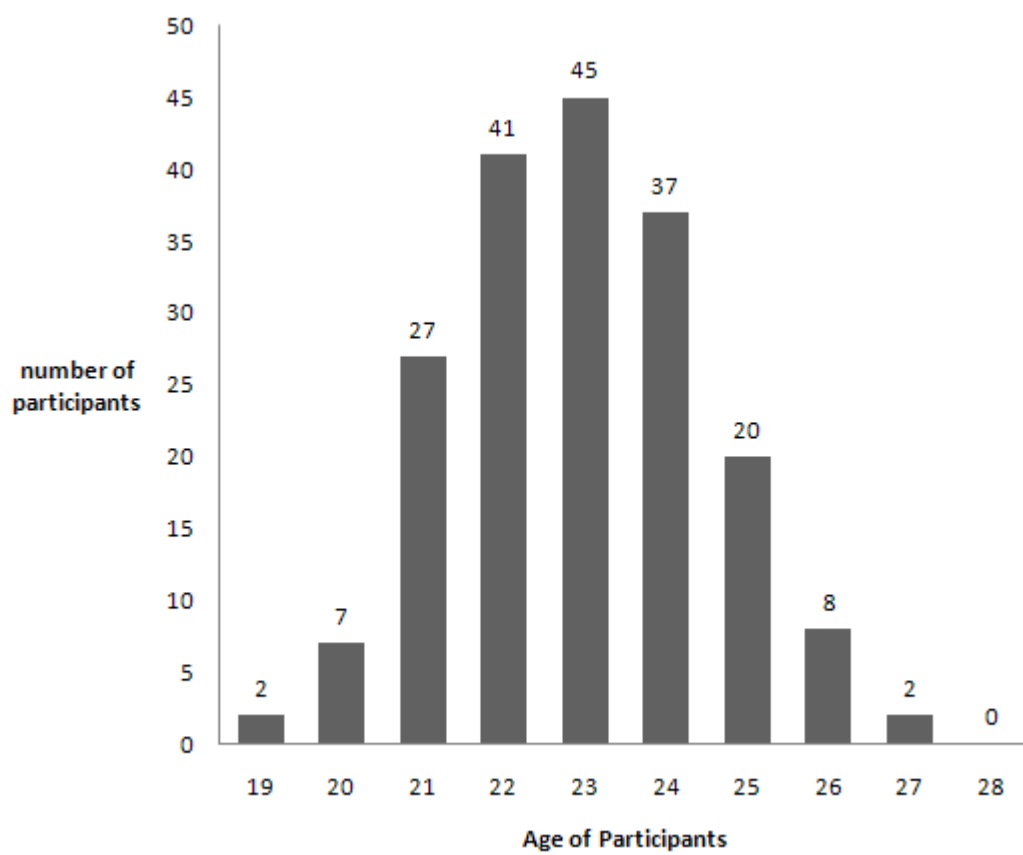


Figure 6.6: Distribution of age (years) of participants.

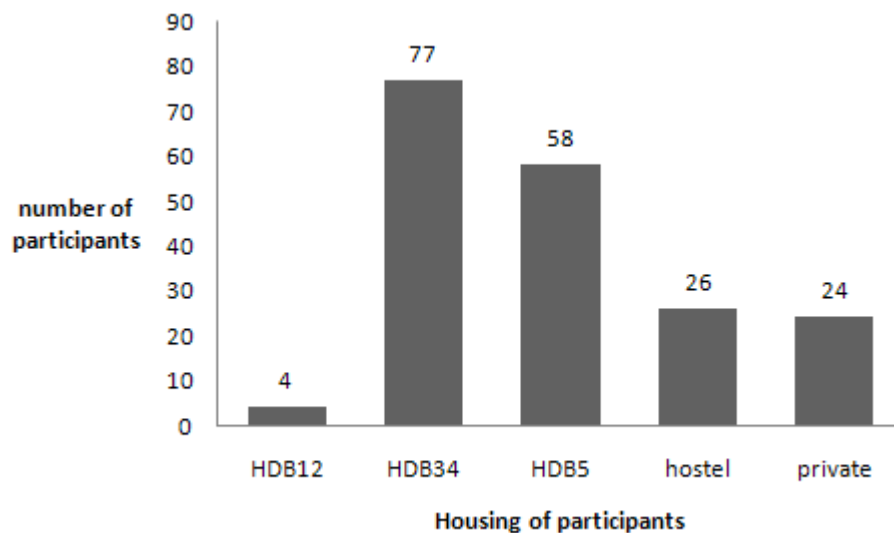


Figure 6.7: Distribution of housing type of participants.

6.4 Learning Beauty

6.4.1 Introduction

Having obtained ratings for 300 tracks, each by roughly 20 participants, the next step is to learn to predict beauty levels from audio accurately.

6.4.2 Method

For each grouping of examples, increasing in number but decreasing in the quality (agreement) of responses, three different feature sets were compared for their ability to classify into one of two categories - beautiful or ugly. The mode of the respondents (the mode is the track more people had rated

beautiful) was considered to be the ground truth or correct answer. The SMO SVM from Weka (see Section 2.4) was used for the machine learning with a polynomial kernel. Leave-one-out cross-validation (see chapter 2) was performed, such that each track was classified based upon information from all the other tracks combined, and the results are given here as the percentage which were correctly classified.

6.4.3 Results

The feature sets are the default MARSYAS set represented by “def”, the default set with chromatic features added: “defchrom”, and lastly all the MARSYAS features. The groupings are: those pairs which were significant at $p < 0.05$ inclusive of the Bonferroni correction (5b), those pairs which were significant at $p < 0.10$ (10b) inclusive of the Bonferroni correction, those pairs which were significant at a simple $p < 0.05$ test (5pc), and finally all the pairs which had a non-tied outcome (all).

	all (298)	5pc (182)	10b (80)	5b (68)
all	67.91%	73.89%	70.51%	74.24%
def	71.96%	78.89%	70.51%	65.15%
defchrom	71.96%	78.89%	70.51%	65.15%

Table 6.2: Performance of different featuresets on different portions of the data, grouped by level of agreement

6.4.4 Discussion

For the two smaller feature sets the best performing grouping is the 5 percent group with 78.89% correct. Having all the examples certainly reduces performance from having the 5% significant ones; this can be seen in every feature row. It is probable that ratings with less significance than this become less and less useful following a law of diminishing returns. The reason for the best performance with the most features being the smallest set of examples is not clear. It could be that more information leads to better classification for a smaller set of examples. It is possible to look at this another way and see that more examples are confounding to a classifier using all the features at its disposal, because as the number of examples increases the distribution becomes more diverse. Perhaps the features are overfitted in the small sample. In each case, however, some predictive power is seen.

6.5 Conclusion

In this chapter a paid survey was conducted spanning 300 extracts of music from around the world and 198 participants. The participants were not particularly diverse but were at least different from what would be readily found here in the UK. Indications are that their ranking of beauty is coloured by their cultural preferences. Learning was performed using their ratings, sectioned into groups by the level of agreement.

The best performing experiment used only default (timbral) features (optionally with chromatic features that did not affect the outcome) from MARSYAS and the 5% set of pairs, which appears to be a sweet-spot in terms of the value of the ratings. Both these combinations achieved 78.89% accuracy. It is interesting that the chromatic features made no difference to the outcome in any of the cases. One reason for this could be that they are tuned to Western scales, and perhaps none of the music, or too little of it, matched up with these tones for them to be informative in classification.

Future work should involve conducting this survey in other countries including in Western countries, to see if the level of agreement varies. The same study with more music would be a valuable expansion, as would different pairings, perhaps from the same country, to see if assessing beauty is easier when comparing music from the same country, and to find out if the same audio features are as useful for predicting ratings. Analysis of the individual track rankings and the “happy” results would also be interesting.

Chapter 7

Beauty and Geography

7.1 Introduction

This chapter tests the hypothesis that the beauty ratings from Singapore were influenced by geographical location of the tracks concerned. First a simple statistical test is run, then a more in-depth attempt to use geographical information to predict beauty ratings, and finally a combination of both audio and geographical features are used for the same prediction task. The responses from the Singapore survey appear to show a preference towards music from around the Singapore region. To test the level of influence of country of origin on the judgements of our raters, first statistical tests were conducted, and then several prediction-by-location experiments were designed which make use of both the absolute position and the relative closeness to Singapore to

predict beauty ratings.

7.1.1 Statistical Dependence on Geography

The standard test for dependence is the chi-squared test. The results must be set into a table such that each combination of beauty, ugly, near and far is represented. For each pair, the track rated more beautiful is considered the beautiful one, and the other is considered ‘ugly’. For each pair, the track nearest to Singapore is considered ‘near’ and the track which is furthest away is considered ‘far’.

A paired Pearson’s chi-squared test was performed on the results at the 5% level (see Table 7.1 and for the “all” group (actually all pairs with a modal result, which happens to be n-1 because only one pair had a completely even set of ratings), found in Table 7.2.

	Near	Far
Beauty	78	13
Far	13	78

Table 7.1: 5% results of χ^2 significance test

	Near	Far
Beauty	107	42
Far	42	107

Table 7.2: 5% results of χ^2 significance test

The Pearson’s chi test statistic is defined in equation (7.1). χ^2 is the test statistic, O_i is the observed frequency, E_i is the expected frequency - in this

case $\frac{1}{4}$ of the total tracks, and n is the number of cells, in this case: 4. The null hypothesis H_0 would be that being nearer has no effect on whether the track is selected as more beautiful by raters.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (7.1)$$

Its value for the “5%” group is 92.86, which when compared the the chi squared distribution is greater than the critical value for $p < 0.001$ at 1 degree of freedom so it is significant at the 0.1% level. (the critical value for this is 10.83). The value for the “all” group is 130, also significant at the 0.1% level.

7.1.2 Distance from Singapore as a Predictor of Beauty Rating

The closer a track’s home country is to Singapore, the more likely our participants are to have rated it ‘beautiful’.

For each track the great-circle distance from Singapore was calculated. This distance, latitude, and longitude are the features of the data. The same SVM algorithm which predicted beauty correctly 79% of the time for the best group is used. The machine task was to predict: for each track, if it was the winner or loser for beauty, in terms of the majority vote of choices made

by participants. Performance was assessed by leave-one-out cross-validation for several different groupings of tracks:

- the 68 tracks significant at the 5% level with the bonferroni correction,
- the 80 tracks inclusive of the above which were significant at the 10% level with the bonferroni correction,
- the 182 tracks inclusive of the above which were significant at the raw 5% level,
- the 298 (all) tracks.

This result can be compared to those achieved with audio content, which was done in a similar fashion. A t-test will determine the statistical significance of any difference.

The tracks were separated into the classes beauty and ugly. The features used were latitude (to 2 d.p.), longitude (to 2 d.p.) and distance from Singapore (to nearest km), firstly all as one vector and then individually. The distance was calculated as a great-circle distance using the spherical law of cosines. It can be seen in (7.2) and (7.3) with ϕ representing latitude and λ representing longitude, d representing distance, r the radius of Earth, and a and b are points. Each of these was based upon the capital city of the track's originating country. These were normalised and passed through the SMO SVM provided in Weka 3.6.8. It implements John Platt's sequential minimal optimisation (SMO) algorithm for training a support vector classi-

fier (or support vector machine, SVM). This was using a polynomial kernel. Leave-one-out cross-validation was performed in all cases.

$$\Delta\hat{\rho} = \arccos(\sin(\phi_a)\sin(\phi_b) + \cos(\phi_a)\cos(\phi_b)\Delta\lambda) \quad (7.2)$$

and then the distance d is:

$$d = r\Delta\hat{\rho} \quad (7.3)$$

7.1.3 Results

Table 7.3 shows the results of the geographical-only study. The four track groups decrease in quality (rater agreement) as they increase in size. The columns show the performance of the SVM with leave-one-out cross-validation using all three geographical features and then each one in turn. The best performance was achieved with the group of 182 tracks using anything except latitude alone.

track group size	all three	just latitude	just longitude	just distance
68	82.35	61.76	82.35	82.35
80	85.00	57.50	85.00	85.00
182	85.71%	53.8%	85.71%	85.71%
298	71.81	52.68	71.81	72.15

Table 7.3: Combinations of geographical features and the SVM performance predicting beauty ratings

As with the beauty prediction with audio features from chapter 6, the 5-percent level has performed best, suggesting again that this is a good balance point compared with other groupings tried of having enough examples of good enough quality to learn something. However, the general performance that is over and above that of audio alone suggests a strong bias in the listeners. This counters the idea that beauty has a strong influence outside of cultural norms. It is possible that all these pieces are too similar in underlying beauty and so other cues were sought by the listeners. Nevertheless it is a disappointment from the perspective of trying to unearth musical beauty from under the heavy influences of personal preference, social influences and wider cultural influences. Distance alone was the best overall measure since it uses the least information and performs equally for the smaller datasets with longitude and with all three geographic pieces of information, whilst for the largest dataset it actually performed better than these. Latitude is clearly the worst indicator, which makes some sense: the tracks were chosen to be as far apart as possible in their pairs, and tracks on opposite sides of the globe could be 180° apart, whilst considering the way the Earth is populated (more people nearer to the equator) they would be unlikely to be further apart in latitude than perhaps 90° , particularly considering that they need to be on land and opposite or nearly opposite another land mass.

7.2 Experiment 2: Location and Audio

7.2.1 Method

Similarly to Experiment 1, a single vector, multi-feature input combining distance information with each of the three sets of audio features is used for the beauty prediction task.

7.2.2 Results

Table 7.4 shows the combination of audio and geographical features. The rows are as in Table 7.3, and the columns are the three audio featuresets already used but each is now combined with the geographic features.

n	def+geo	defchrom+geo	all+geo
68	72.73%	72.73%	75.76%
80	78.21%	78.21%	75.64%
182	82.78%	82.78%	79.44%
298	72.97	72.97%	73.65%

Table 7.4: Combinations of audio and geographical features and their performance at predicting beauty

Clearly yet again the 5 percent row yields the best results. Despite this, the capabilities of the algorithms are somehow poorer for the combination, perhaps indicating a limitation in the way SVMs are able to prioritise features. It seems that using all the audio features is also still a confounding factor for all tasks but the largest set. Again, the addition of chromatic features leave

the performance unchanged. This is interesting in the light of the findings of Gomez *et al.* who discovered latitude to be correlated with tonal, particularly chromatic, features, whereas longitude was correlated with timbral features [Gomez et al., 2009]. It is possible our dataset is simply not interestingly distributed for latitude. Longitude being correlated with timbral features could be a function of the different instrumentation being detected by these. Instruments vary with longitude simply because instruments vary with distance. These results in combination with Section 7.1.2 and Chapter 6 seems to have confirmed their findings since longitude was important and timbral features were important in predicting beauty, and latitude and chroma have separately proved to add little value in predicting beauty, which in Section 7.1.1 statistically proven to be strongly linked to geographic location.

7.3 Conclusion

Several different tests including statistical tests and learning tasks were conducted to assess the correlation between the closeness of a track to Singapore and it being rated more beautiful. In all tests the correlation was shown to be present. It is evident that the Singapore participants were swayed by some geographically related influence, which after conducting these tests is only more clear. In doing so a clear method for comparing the influences of different feature sets has been developed which gives more insight than a

simpler statistical test. It has also been shown that, for this dataset at least, statistical agreement at the 5% level is a good basis for learning. The best performance combining both audio and geographical information was with the default or default and chromatic features, attaining 82.78% accuracy. The best performing result using geographic information alone was achieved with any of longitude, distance from Singapore, or latitude, longitude and distance from Singapore in combination, showing only latitude to be almost worthless in predicting beauty. The correlation between being the nearest track and being rated most beautiful was significant at the 0.1% level for the 5% group. Latitude correlation with chromatic factors and longitude being correlated with timbral factors is supported by these results. Further work would entail surveying several different populations with the same music to see what the differences are in perception and how strongly they are related to location, rather than audio features, in different places. It would be interesting to investigate this variance. It would also be useful to repeat the experiment with tracks that are instead both from the same country in each pair to see if this makes the task easier or harder. Other machine learning algorithms such as those reviewed in Chapter2 could be tried to validate the results of this study.

Chapter 8

Conclusion

8.1 Summary of Work

The purpose of this work was to investigate what is meant by 'beauty' in music, if it relates to audio features and what other factors come into play. The use of non-Western music and the influence of geography in the perception of beauty were important aspects considered. There has been a lot of previous work in analysing Western music with computational approaches on audio features, but little covering beauty or non-Western music.

Different representations for audio features and different machine learning algorithms for learning on audio features were tested, with MARSYAS all features and Support Vector Machines performing well. Predicting beauty was first investigated with a small Last.fm dataset and later with a larger

world music survey with Singaporean participants. Predicting the geographical location of world music was attempted, producing some promising results and the development of interesting methods for accommodating the surface of Earth in doing so. The Singaporean beauty ratings were predicted from audio content, geographic content and a combination of both, showing strong correlations between longitude, distance, and timbral features with the beauty ratings, which were statistically very closely linked with distance from Singapore. Beauty in music as rated by our participants was culturally related and timbre is a good pointer to cultural differences.

The work is a valuable addition to the field of Computational Musicology. We have gained new insights into the possibility of detecting beauty in audio signals. Predicting the geographical location of world music was shown to be possible albeit at an early stage, and some of the findings of Gomez *et al.* regarding important audio features for geographic prediction were confirmed [Gomez et al., 2009]. Cultural influence on the rating of music was shown to affect the concept of beauty in music, at least as understood by our survey participants. It was also demonstrated that these ratings were more easily predicted by geographical information than by audio features.

8.2 Initial Experiments

Methods for obtaining ground truth from raters and from benchmark datasets were tested and found to be feasible. Machine learning algorithms were compared on the benchmark data set and SVMs were amongst the best-performing. Beauty was predicted on a small selection of Last.fm tags with good accuracy. Two different feature sets from MARSYAS were compared and the larger was found to improve the results when more complex learning algorithms were used (such as the SVM). This set of experiments provided insights into the most useful feature representation, machine learning algorithms, survey techniques and possibility of detecting beauty by machine.

8.3 Geographic Prediction

The distribution of music from around the world was investigated. The problem was cast as training a machine learning program to predict the geographical origin of pieces of music. World music was described using different sets of standard audio descriptors from MARSYAS. To predict the location of origin of the music several methods were developed, designed to deal with the spherical surface topology based upon a modified k -Nearest-Neighbour algorithm. The utility of *a priori* geographical knowledge in the predictions was investigated in the form of a land and sea mask, and a population distribu-

tion overlay. Results were statistically significant and encouraging but with room for improvement. This was a contribution to the field of geographical ethnomusicology.

8.4 Singapore Survey

In this chapter a paid survey was conducted with Singaporean participants on world music. Statistical analysis indicated that their ranking of beauty was coloured by their cultural preferences. Learning was performed using their ratings, sectioned into groups by the level of agreement. It was proven that beauty ratings can be predicted from audio features. The quality of ratings required for best prediction was discovered. Chromatic features were seen to make no difference in prediction.

8.5 Geographical Influence on Ratings

Several different tests including statistical tests and learning tasks were conducted to assess the correlation between the closeness of a track to Singapore and it being rated more beautiful. In all tests the correlation was shown to be present. It is evident that the Singapore participants were swayed by some geographically related influence, which after conducting these tests is only more clear. In doing so a clear method for comparing the influences

of different feature sets has been developed which gives more insight than a simpler statistical test. It has also been shown that, for this dataset at least, statistical agreement at the 5% level is a good basis for learning. Using audio and geographic information performed worse than using just geographic information. Longitude being correlated with timbral factors is supported by these results, which gives weight to similar results found by others on the field. This may be as a result of instrumentation detection.

8.6 Future Work

An expansion of the Last.fm to a much larger dataset would be useful to see how those example which are less agreed-upon as being beautiful confound or contribute to prediction. An individual consideration of the contribution of each MARSYAS feature would be helpful to determine with more fine-graining what representation is best for beauty and for geography. Extra features would also be of interest, such as the fine chromatic feature used by Gomez *et al.* [Gomez and Herrera, 2008]. A larger world music corpus with both more tracks from each country, and more countries represented, would improve prediction results and enrich any future surveys. It would be better to have access to the exact location of origin of the music, rather than just the capital or population centroid, as most countries have strong regional variations in style. Some cultures change drastically over small areas, some are unchanged over large expanses, and this could also be learnt.

It would also be useful to explore the possibility of filtering and pre-selection of descriptors for both beauty and geographic experiments. Many other forms of machine learning could be applied: neural-networks, support vector machines, decision trees, etc. Here there was only time to apply the most promising technique to all the different datasets. Conducting world music surveys in other countries including in Western countries, would be useful to see if the level of agreement varies. The same study with more music would be a valuable expansion, as would different pairings, perhaps from the same country, to see if assessing beauty is easier when comparing music from the same country, and to find out if the same audio features are as useful for predicting ratings. Analysis of the individual track rankings and the “happy” results would also be interesting for comparison. It would be worth investigating to see what the differences are in perception and how strongly they are related to location, rather than audio features, in different places. It would also be useful to repeat the experiment with tracks in different pairings.

8.7 Remarks

Computational Musicology is a new field and this thesis contributes to computational understanding of beauty in music, geographical influences in music, geographical influences in appreciation of music, and techniques for representing music across the globe. There are many threads of interest to be

followed up from the work and it is hoped that others will do so. Music Information Retrieval is growing in popularity and in commercial value. It has ties with speech recognition techniques and other applications that begin with entertainment recommendations and end with deep investigations of what makes us like music in the first place, and so, in part, what makes us human.

Appendix A

Full Results from Singapore Survey

The table headings “beauty” and “happy” refer to the percentage agreement between raters. Column “n” is the number of raters, as this varied slightly and affects the statistical calculation. $p(h_0)b$ and $p(h_0)h$ are the respective probabilities of the null hypothesis, that is that there is no difference in the beauty or happiness ratings on a particular pair. “winner” and “loser” are the winning and losing tracks in each pair, that is, the winner is the one rated most beautiful. The beginning of each track name is the 2-letter ISO-3166-1 country code for that track, so the countries of origin can be fairly easily observed.

Table A.1: Agreement about beauty and happiness of non-Western music amongst Singaporean listeners

beauty	happy	n	$p(h_0)b$	$p(h_0)h$	winner	loser
100%	100%	20	0.00000	0.00000	MA_29_16BintBladi	AU_83_02KurongkBoy
100%	85%	20	0.00000	0.00258	JP_84_08HiyamiKach	BR_78_13Linda
100%	75%	20	0.00000	0.04139	TH_16_01HaeNangMae	CV_80_02ValsaAzul
100%	100%	20	0.00000	0.00000	TH_16_13SiengSoong	CV_79_12Rabecadai
100%	95%	20	0.00000	0.00004	CN_19_02LeRendez	CV_80_04Odjudagu
100%	95%	19	0.00000	0.00008	TW_22_15YiNaNaZao	BR_77_02Beija
100%	100%	19	0.00000	0.00000	TW_22_03PaYaSiVINo	BR_78_02TakeSarava
100%	95%	19	0.00000	0.00008	TW_22_19LaNuKoDan	BR_3_07NoPrego
100%	100%	19	0.00000	0.00000	AU_83_08DragNet	MA_6_12MinYoumi_SinceAlways_
100%	100%	17	0.00002	0.00002	CN_19_01LeVentDams	CV_79_10Chap
100%	59%	17	0.00002	0.62906	SN_4_03ToubaDaruSalaam	CN_21_17TheNightOf
95%	80%	20	0.00004	0.01182	AU_83_11Nyindi	MA_33_05Touria
95%	90%	20	0.00004	0.00040	TH_18_03CherdNawk	CV_80_01Dordiamor
95%	100%	20	0.00004	0.00000	TH_18_01ChatriOver	CV_79_06AfricaUmDi
95%	95%	20	0.00004	0.00004	TH_16_06AnElephant	CV_3_07MonteCara
95%	90%	20	0.00004	0.00040	TH_17_02TheFloatin	CV_80_11Col
95%	100%	20	0.00004	0.00000	TH_17_01Soundsoft	CV_79_05Toy
95%	65%	20	0.00004	0.26318	TH_16_07MooLamLamT	CV_80_14Cretcheu
95%	50%	20	0.00004	1.00000	TH_16_18SaoNaSangF	CV_80_08Lamentodeumemigrante
95%	100%	20	0.00004	0.00000	TH_16_17SaoNoongSa	CV_80_13Equilibrio
95%	95%	20	0.00004	0.00004	CN_5_12InTheSettingOfTheSun	CV_80_10Landudiamo
Continued on next page						

Table A.1 – continued from previous page

beauty	happy	n	p(h0)b	p(h0)h	winner	loser
95%	100%	19	0.00008	0.00000	TW_22_17SeNaiKiDaL	BR_77_03MulateDoBunde
95%	100%	19	0.00008	0.00000	TW_22_11Lalai	BR_4_04A ganju
95%	100%	19	0.00008	0.00000	TW_22_22ChantDeLou	BR_2_10TiveRazao_IWasRight_
95%	58%	19	0.00008	0.64761	MA_29_06Sana	AU_83_10SikO
95%	100%	19	0.00008	0.00000	AU_83_05WonggaInit	MA_87_01AMueyAMuey
95%	84%	19	0.00008	0.00443	AU_83_01Saltwater	MA_33_04LeilaLill
95%	95%	19	0.00008	0.00008	MA_29_10Sana	AU_83_03NativeBorn
95%	100%	19	0.00008	0.00000	AU_83_13Bushfire	MA_33_07YedidimHio
90%	52%	21	0.00022	0.83181	JM_27_02JahShowThe	ID_26_15BengawanSo
94%	100%	17	0.00027	0.00002	TW_22_20Balen	BR_3_11VelhaInfancia
94%	88%	17	0.00027	0.00235	TW_22_23ChantDEnfa	BR_3_16CucurucucuPaloma
94%	82%	17	0.00027	0.01273	CN_21_07HongNiangH	GM_45_05Bitillo
94%	65%	17	0.00027	0.33231	GM_45_12Aminatta	CN_21_08ShanxiAi
90%	90%	20	0.00040	0.00040	AU_83_09HeartOfMyP	MA_35_05SabaAtuRijal
90%	95%	20	0.00040	0.00004	AU_83_12Celebratio	MA_33_03Bay
90%	95%	20	0.00040	0.00004	JP_84_13AsadoyaYun	BR_78_12Curiosidad
90%	65%	20	0.00040	0.26318	KH_92_02TrachToch	BZ_85_08SinPrecioW
90%	95%	20	0.00040	0.00004	KH_92_14NeangMeo	BZ_85_10AyuhaNiduu
90%	90%	20	0.00040	0.00040	CN_20_05PuAnZhouSu	CV_79_07NhaFiDJo
89%	100%	19	0.00073	0.00000	TW_22_14SeNaiKiNia	BR_78_14HempeFesta
89%	84%	19	0.00073	0.00443	TW_22_25ChantDAdie	BR_78_10MeninodoPe
89%	100%	19	0.00073	0.00000	KH_92_09TrapeangPa	CU_3_06HayQueEntrarleAPalosAEse
89%	94%	18	0.00131	0.00014	TW_22_12MaoWaPaPaL	BR_78_11LifeGods
89%	89%	18	0.00131	0.00131	TW_22_01EHoYi	BR_78_03SwingdaCor

Continued on next page

Table A.1 – continued from previous page

beauty	happy	n	p(h0)b	p(h0)h	winner	loser
86%	100%	21	0.00149	0.00000	KH_92_08KlangChana	PE_6_03SeMeVanLosPies
88%	100%	17	0.00235	0.00002	TW_22_24ChantAlter	BR_78_16ExuAnan
88%	100%	17	0.00235	0.00002	TW_22_16BuSiLiVoLo	BR_78_09CoracaodeB
88%	88%	17	0.00235	0.00235	CN_20_04DeshengLin	SN_45_10DiamanoBif
85%	80%	20	0.00258	0.01182	AU_83_06Bullima	MA_29_01Mashaliya
85%	90%	20	0.00258	0.00040	KH_92_11Somplrov	BZ_85_03Miami
85%	90%	20	0.00258	0.00040	KH_92_10Lam	BZ_85_02WeyuLaarig
85%	70%	20	0.00258	0.11532	CV_79_14CPLP	TH_16_05RangJaiRaiWan
85%	65%	20	0.00258	0.26318	TH_16_12TheNangHon	CV_79_01DorDiAmor
85%	75%	20	0.00258	0.04139	TH_18_04PhlengReua	CV_79_04PapaJoachi
85%	90%	20	0.00258	0.00040	TH_16_15TamHaKujee	CV_2_12MundoENos
85%	70%	20	0.00258	0.11532	TH_16_10RoopKhaoKr	CV_80_07Tchap
85%	65%	20	0.00258	0.26318	TH_16_08AmazingIsa	CV_79_15CorDiRosa
84%	89%	19	0.00443	0.00073	TW_22_09LauAhLuMeD	BR_3_01SantaMassaChegou
84%	95%	19	0.00443	0.00008	ID_26_11JogedLaksm	PR_1_05P rakatun
84%	95%	19	0.00443	0.00008	TM_65_04FromTheSta	MX_5_10LaCumbiaDeIMole
84%	74%	19	0.00443	0.06357	GN_6_10Sabou	JP_84_12Ubue
84%	53%	19	0.00443	1.00000	JP_84_05TakiosSora	GN_59_13Damensna
84%	100%	19	0.00443	0.00000	TH_16_04LerkDaiLer	US_5_17BetterWayWarMix
81%	57%	21	0.00720	0.66362	JM_27_11CrashieFir	ID_26_13Kucap
80%	90%	20	0.01182	0.00040	AU_83_14Black	MA_29_15InsirafFro
80%	90%	20	0.01182	0.00040	JP_84_01Makura	BR_77_10An
80%	80%	20	0.01182	0.01182	JP_84_14JamesBondT	BR_78_06GingadaBale
80%	90%	20	0.01182	0.00040	KH_92_03KamanPrath	BZ_85_04Baba

Continued on next page

Table A.1 – continued from previous page

beauty	happy	n	p(h0)b	p(h0)h	winner	loser
80%	100%	20	0.01182	0.00000	KH_92_12Lam	BZ_85_07BeiBaGoAwa
80%	80%	20	0.01182	0.01182	KH_92_05KravnayCho	BZ_85_06Gagaabadib
80%	80%	20	0.01182	0.01182	KH_92_06Pyadeun	BZ_85_01Watina
80%	85%	20	0.01182	0.00258	TM_65_13Gongurbash	MX_6_09SaleSobrand
80%	95%	20	0.01182	0.00004	TH_16_16SoDuangThe	CV_80_06Homi
80%	90%	20	0.01182	0.00040	TH_17_04Heartofthe	CV_80_03GuerraFria
82%	94%	17	0.01273	0.00027	CN_21_11LanGueiJi	GM_45_09KairabaJabi
82%	82%	17	0.01273	0.01273	CN_20_03ChoudianZh	SN_35_08Dembe
79%	53%	19	0.01921	0.82380	JM_27_05DreadCalle	ID_26_05JerukManis
79%	74%	19	0.01921	0.06357	ZA_4_05LahlUmlenze	JP_84_15HaisaiOjis
79%	84%	19	0.01921	0.00443	JP_84_09MojiBanana	GN_4_07Wawata
79%	84%	19	0.01921	0.00443	GN_59_04LanNaya	JP_84_02KakinOndo
79%	84%	19	0.01921	0.00443	JP_84_10Utuwaskara	GN_59_09Haidara
76%	100%	21	0.02660	0.00000	CN_19_04AffleeEst	AR_4_09TierraColoradaRedLand
76%	81%	21	0.02660	0.00720	JM_27_07TheGeneral	ID_26_03KaretaMala
78%	100%	18	0.03088	0.00001	TW_22_02EYao	BR_77_06SeVocSeFor
75%	80%	20	0.04139	0.01182	KH_92_13Lam	BZ_85_09YaganeMyCa
75%	75%	20	0.04139	0.04139	KH_92_04LarSmeurRo	BZ_85_11Ayo
75%	80%	20	0.04139	0.01182	TH_16_03PhinSoloTr	CV_80_15NhaRiqueza
75%	75%	20	0.04139	0.04139	TH_16_14KatikarHua	CV_79_11Filosofia
75%	80%	20	0.04139	0.01182	TH_16_09OhOhOh	CV_79_17DorDiNhAlm
76%	76%	17	0.04904	0.04904	CN_20_08XiaoKaimen	CV_79_08Cornologia
74%	63%	19	0.06357	0.35928	JM_27_10SoLongRast	ID_25_03Gending
74%	63%	19	0.06357	0.35928	JM_27_22NattyB	ID_26_09Sumbawa

Continued on next page

Table A.1 – continued from previous page

beauty	happy	n	p(h0)b	p(h0)h	winner	loser
74%	84%	19	0.06357	0.00443	ID_26_06DarDerDor	JM_27_13Repatriati
74%	68%	19	0.06357	0.16707	JM_27_08DaylightSa	ID_26_04Begadang
74%	95%	19	0.06357	0.00008	TW_22_13OuiNaRuWan	BR_77_05ToqueDeTim
74%	89%	19	0.06357	0.00073	TH_18_02SathukarnT	CU_1_01ElCapitan
74%	74%	19	0.06357	0.06357	GN_59_05Bassa	JP_84_17HaNaMi
71%	95%	21	0.07835	0.00002	ID_26_08CeurikRahw	CO_5_01LagartijaAzul
71%	71%	21	0.07835	0.07835	JM_27_17Thirst	ID_26_01Sambasunda
70%	95%	20	0.11532	0.00004	BR_78_04AVidaBoa	JP_84_07MangetsuNo
70%	55%	20	0.11532	0.82380	TH_16_11MaKorThoTa	CV_79_09GritoMagoa
70%	65%	20	0.11532	0.26318	TH_16_19PongLangEn	CV_79_16Sarapilh
70%	50%	20	0.11532	1.00000	CV_80_12Ra	CN_21_15ChiangWeiCuCuKai
70%	90%	20	0.11532	0.00040	CN_20_05LiuyaoJin	CV_79_03Dan
71%	94%	17	0.14346	0.00027	CN_19_05LeProceesK	GM_45_06Salimata
71%	59%	17	0.14346	0.62906	CN_21_09YellowBana	SN_45_03Loodo
71%	82%	17	0.14346	0.01273	CN_21_13TirikBosta	SN_45_07Letter
68%	89%	19	0.16707	0.00073	KH_92_01Sathouka	CU_6_11Chiquichaca
68%	95%	19	0.16707	0.00008	TH_17_03ASarlitNi	MX_4_06Viborita
67%	57%	21	0.18925	0.66362	JM_27_16CupOfTea	ID_24_06LegongKrat
67%	100%	18	0.23788	0.00001	BR_77_01BraseiraAr	TW_22_05MiYoMe
65%	75%	20	0.26318	0.04139	MA_29_14Mawal	AU_83_07KavaSong
65%	50%	20	0.26318	1.00000	BR_77_04CantoProMa	JP_84_19ShiChome
65%	100%	20	0.26318	0.00000	KH_92_07SompougnKl	BZ_85_05LidanAban
65%	100%	17	0.33231	0.00002	TW_22_07NaKuMo	BR_78_08ComMuzenza
65%	94%	17	0.33231	0.00027	CV_80_09Simentera	CN_21_02Silaihwar

Continued on next page

Table A.1 – continued from previous page

beauty	happy	n	p(h0)b	p(h0)h	winner	loser
65%	76%	17	0.33231	0.04904	GN_59_08Sara70	CN_21_10NocturnalLight
63%	63%	19	0.35928	0.35928	JM_27_09ItalFighti	ID_24_03TabuhPisan
63%	84%	19	0.35928	0.00443	ID_24_05SinomLadra	JM_27_18ISawESaw
63%	95%	19	0.35928	0.00008	TW_22_10OuHaLaYiYo	BR_78_05PitadadeTa
62%	86%	21	0.38331	0.00149	CI_2_11Abiani	TV_6_11Alamagoto
62%	90%	21	0.38331	0.00022	ID_26_02AnomanObon	JM_27_03Cb200
60%	90%	20	0.50344	0.00040	BR_77_09Ai	JP_84_11NyoraiShiz
60%	75%	20	0.50344	0.04139	BR_77_14MargaridaP	JP_84_06AmagoiBushi
60%	80%	20	0.50344	0.01182	CV_80_05TributoaSa	CN_20_09TaiHuaJiao
59%	100%	17	0.62906	0.00002	TW_22_04ToYiSo	BR_77_08MimarVoc
59%	94%	17	0.62906	0.00027	TW_22_08Maluppallim	BR_77_11ALatinha
59%	88%	17	0.62906	0.00235	CN_21_01YiWuSouYou	CV_79_02FalsoTeste
58%	68%	19	0.64761	0.16707	ID_26_14BolehBersu	JM_27_21WarIsOver
58%	74%	19	0.64761	0.06357	ZA_5_02KuraUone	JP_84_04NikataBush
57%	67%	21	0.66362	0.18925	JM_27_01Ragnampiza	ID_24_01TabuhGari
57%	81%	21	0.66362	0.00720	ID_25_01Ketawang	JM_27_15MarijuanaI
57%	62%	21	0.66362	0.38331	JM_27_04CokaneInMy	ID_24_02Gambang
53%	94%	17	0.81453	0.00027	TW_22_18LaLiZoKo	BR_77_13Pap
56%	89%	18	0.81453	0.00131	TW_22_21ChantDeReu	BR_77_12
53%	79%	19	0.82380	0.01921	JM_27_19Plantation	ID_25_04Bubaran
53%	68%	19	0.82380	0.16707	JM_27_06FlatFootHu	ID_26_10Rentak106
55%	75%	20	0.82380	0.04139	BR_77_07FricoteDaT	JP_84_18Utag
55%	85%	20	0.82380	0.00258	BR_78_07DeusadoEba	JP_84_16AgariJo
55%	75%	20	0.82380	0.04139	JP_84_03FukkoBushi	BR_78_15MamaeQueri

Continued on next page

Table A.1 – continued from previous page

beauty	happy	n	p(h0)b	p(h0)h	winner	loser
55%	95%	20	0.82380	0.00004	CV_79_13FundoBaxo	TH_16_02LamYaiLamP
52%	81%	21	0.83181	0.00720	JM_27_20MeltingPot	ID_25_02Gending
50%	94%	18	1.00000	0.00014	No clear	winner
52%	86%	21	1.00000	0.00149	ID_26_07LosQuinTal	JM_27_14Cornbread
53%	89%	19	1.00000	0.00073	ID_24_04Barong	JM_27_12BionicDrea
53%	84%	19	1.00000	0.00443	ID_26_12PageSakari	BZ_85_12
53%	63%	19	1.00000	0.35928	MA_29_17HabibaUJar	AU_83_04TjapukaiSu
53%	79%	19	1.00000	0.01921	MX_6_07Sr.Judas	TM_65_14SongOfKark
53%	88%	17	1.00000	0.00235	CN_21_12TheGreenBr	SN_46_03Ndiawolou

Appendix B

User Demographic Data for the Singapore Survey

Table B.1: Demographics of Listeners

group	age	gender	race	housing	nationality	first language	second language
0	27	M	Chinese	HDB34	singaporean	English	Mandarin
0	24	M	Indian	HDB34	Singaporean	English	Malay
0	23	M	Chinese	HDB34	Singaporean	English	Mandarin
0	24	M	Punjabi	HDB5	Singaporean	English	Malay
0	24	M	Chinese	HDB34	Singaporean	English	Mandarin
0	24	M	Chinese	HDB34	Singaporean	English	Mandarin
0	26	M	Chinese	HDB34	singaporean	English	Mandarin
0	25	M	Chinese	HDB34	Singapore	English	Mandarin
0	22	F	Chinese	private	Singaporean	English	Mandarin
0	23	M	Chinese	HDB34	Singaporean	English	Mandarin
0	21	F	Chinese	HDB34	Singapore Citizen	English	Mandarin
0	23	M	Chinese	private	Singaporean	Mandarin	English
0	23	F	Chinese	private	Singaporean	English	Mandarin
0	22	F	Malay	HDB5	Singaporean	English	Malay
0	24	F	Chinese	HDB34	Singaporean	English	Mandarin
0	23	M	Chinese	HDB34	Singaporean	English	Mandarin
0	21	F	Chinese	HDB5	singaporean	Mandarin	English
0	24	M	Chinese	HDB34	Singaporean	English	Mandarin
0	22	M	Chinese	HDB34	Indonesian	English	Mandarin
0	22	F	Chinese	HDB5	Singaporean	Mandarin	English
0	24	M	Chinese	HDB34	Singaporean	English	Mandarin
1	23	F	Chinese	HDB5	Singaporean	English	Mandarin
Continued on next page							

Table B.1 – continued from previous page

group	age	gender	race	housing	nationality	first language	second language
1	24	M	Chinese	hostel	Malaysian	English	Mandarin
1	26	M	Chinese	HDB34	Singaporean	Mandarin	English
1	27	M	Chinese	HDB5	SINGAPOREAN	English	Mandarin
1	23	F	Chinese	HDB34	Singaporean	English	Mandarin
1	23	M	Chinese	private	Singaporean	English	Mandarin
1	25	M	Indian	HDB34	Singaporean	English	Tamil
1	24	M	Chinese	HDB34	singaporean	English	Mandarin
1	25	M	Chinese	HDB34	Singaporean	English	Mandarin
1	23	M	Chinese	private	singaporean	English	Mandarin
1	23	F	Chinese	HDB34	Singaporean	English	Mandarin
1	23	F	Chinese	HDB34	Singaporean	English	Mandarin
1	22	F	Chinese	HDB5	Singaporean	English	Mandarin
1	25	M	Chinese	HDB34	Singaporean	English	Mandarin
1	24	M	Chinese	HDB5	singaporean	English	Mandarin
1	22	F	Chinese	HDB34	Singaporean	English	Mandarin
1	23	F	Chinese	private	Singaporean	English	Mandarin
1	21	F	Chinese	HDB5	Singaporean	Mandarin	English
1	23	M	Chinese	HDB5	Singaporean	Mandarin	English
2	24	M	Chinese	HDB5	Singaporean	English	Mandarin
2	22	F	Chinese	HDB34	Singaporean	Mandarin	English
2	0		NULL	NULL	NULL	NULL	NULL
2	25	M	Chinese	HDB34	Singaporean	English	Mandarin
2	22	F	Chinese	HDB5	Singaporean	English	Mandarin
2	23	F	Chinese	private	Singapore	English	Mandarin

Continued on next page

Table B.1 – continued from previous page

group	age	gender	race	housing	nationality	first language	second language
2	24	M	Chinese	HDB34	SINGAPOREAN	Mandarin	English
2	24	M	Chinese	HDB5	Singaporean	English	Mandarin
2	22	F	Chinese	HDB34	Singaporean	Mandarin	English
2	25	M	Chinese	private	Singaporean	English	Mandarin
2	21	F	Chinese	hostel	Indonesian	Bahasa Indonesia	English
2	20	F	Chinese	HDB5	Singapore Citizen	English	Mandarin
2	22	M	Chinese	hostel	Malaysian	Mandarin	English
2	21	F	Chinese	private	Singaporean	English	Mandarin
2	23	F	Chinese	private	Singaporean	English	Mandarin
2	25	M	Chinese	HDB5	Singaporean	English	Mandarin
2	23	M	Chinese	HDB5	Singaporean	English	Mandarin
2	21	F	Chinese	HDB5	Singaporean	English	Mandarin
2	22	F	Indian	HDB5	Indian (SPR)	Tamil	English
2	20	F	Chinese	HDB5	Singaporean	Mandarin	English
3	0		NULL	NULL	NULL	NULL	NULL
3	26	M	Chinese	HDB34	Singaporean	English	Mandarin
3	22	M	Chinese	HDB5	singaporean	English	Malay
3	20	F	Malay	HDB5	SINGAPOREAN	English	Malay
3	22	F	Chinese	HDB34	Singaporean	English	Mandarin
3	25	M	Chinese	HDB34	Singaporean	English	Mandarin
3	22	F	Chinese	hostel	Malaysia	Mandarin	English
3	21	F	Chinese	HDB5	Singaporean	English	Mandarin
3	20	F	Chinese	HDB5	Malaysian (S'pore PR)	Mandarin	English
3	23	F	Chinese	private	singaporean	English	Mandarin

Continued on next page

Table B.1 – continued from previous page

group	age	gender	race	housing	nationality	first language	second language
3	24	M	Chinese	private	Singaporean	English	Mandarin
3	25	M	Chinese	HDB5	Singaporean	English	Mandarin
3	25	M	Chinese	HDB34	Singapore	Mandarin	English
3	22	F	Malay	HDB5	Singapore	English	Malay
3	23	F	Chinese	HDB12	Indonesian	Indonesian	English
3	22	F	Chinese	HDB5	Singaporean	English	Mandarin
3	22	M	Chinese	hostel	Indonesian	Bahasa	English
3	23	F	Chinese	HDB34	Singaporean	English	Mandarin
3	22	F	Chinese	HDB34	Singaporean	English	Mandarin
3	26	M	Chinese	HDB34	CHINA	Mandarin	English
4	22	M	Chinese	HDB12	Chinese	Mandarin	English
4	25	M	Javanese	HDB34	Singaporean	English	Malay
4	24	F	Chinese	HDB34	Singaporean	Mandarin	English
4	21	F	Chinese	HDB5	Singaporean	English	Mandarin
4	23	F	Chinese	hostel	Malaysian	Mandarin	English
4	20	F	Chinese	private	Singaporean	English	Mandarin
4	25	M	Chinese	HDB5	Singapore	English	Mandarin
4	21	F	Chinese	private	Singaporean	English	Mandarin
4	23	F	Chinese	HDB34	Singaporean	English	Mandarin
4	21	F	Chinese	HDB34	singaporean	English	Mandarin
4	24	M	Chinese	HDB34	Singaporean	English	Mandarin
4	22	F	Chinese	HDB5	Singaporean	Mandarin	English
4	22	F	Chinese	HDB34	Singaporean	English	Mandarin
4	23	M	Eurasian	HDB34	Kazakh	Kazakh	Russian

Continued on next page

Table B.1 – continued from previous page

group	age	gender	race	housing	nationality	first language	second language
4	20	F	Chinese	HDB5	Singaporean	English	Mandarin
4	25	M	Chinese	HDB5	Singaporean	English	Mandarin
4	22	M	Chinese	HDB5	Singaporean	English	Mandarin
4	23	M	Chinese	HDB34	Singaporean	English	Mandarin
4	21	F	Chinese	HDB34	Singaporean	English	Mandarin
4	23	M	Chinese	HDB34	Singaporean	English	Mandarin
5	21	M	javanese	HDB34	indonesian	bahasa indonesia	English
5	21	F	Pakistani	HDB5	Pakistani	English	Urdu
5	19	F	Chinese	HDB34	Singaporean	English	Mandarin
5	25	M	Chinese	HDB5	Sporean	English	Mandarin
5	25	M	Chinese	private	Singaporean	English	Mandarin
5	23	M	Chinese	HDB12	malaysian	Mandarin	English
5	21	F	Chinese	HDB34	Singaporean	Mandarin	English
5	23	M	Chinese	hostel	Malaysian	Mandarin	English
5	23	M	Chinese	HDB5	Singaporean	English	Mandarin
5	24	M	Chinese	HDB5	Singaporean	English	Mandarin
5	21	F	Chinese	HDB34	Singaporean	English	Mandarin
5	22	F	Indian	hostel	India	English	Hindi
5	23	M	Chinese	HDB34	singaporean	Mandarin	English
5	21	F	Chinese	HDB34	SG	English	Mandarin
5	22	F	Chinese	HDB34	SINGAPOREAN	Mandarin	English
5	21	F	Chinese	HDB34	Singaporean	English	Mandarin
5	25	M	Chinese	HDB34	Singaporean	Mandarin	English
5	24	M	Chinese	HDB5	Singapore	English	Mandarin

Continued on next page

Table B.1 – continued from previous page

group	age	gender	race	housing	nationality	first language	second language
5	22	F	Chinese	hostel	malaysian	Mandarin	English
5	22	F	Chinese	hostel	Chinese	Mandarin	English
6	23	F	Chinese	hostel	malaysian	Mandarin	English
6	26	M	Chinese	HDB5	Singaporean	English	Mandarin
6	23	F	Chinese	HDB34	Singaporean	English	Mandarin
6	20	F	Chinese	HDB34	HONG KONG	English	Mandarin
6	22	F	Chinese	HDB5	Singaporean	English	Mandarin
6	22	F	Chinese	HDB34	Singaporean	English	Mandarin
6	24	F	Chinese	hostel	Malaysian	English	Mandarin
6	26	M	Chinese	HDB34	Singaporean	Mandarin	English
6	23	M	Chinese	hostel	Chinese	Mandarin	English
6	22	F	Chinese	HDB34	Singaporean	English	Mandarin
6	24	M	Chinese	hostel	Malaysian	Mandarin	English
6	24	M	Chinese	hostel	MALAYSIAN	Mandarin	English
6	25	M	Chinese	private	Singaporean	English	Mandarin
6	24	M	Chinese	private	Singaporean	English	Mandarin
6	25	M	Chinese	HDB5	SINGAPOREAN	English	Mandarin
6	26	M	Chinese	HDB34	Singaporean	Mandarin	English
6	23	M	Chinese	hostel	China	Mandarin	English
6	26	M	Chinese	HDB5	Singaporean	English	Mandarin
6	21	F	Chinese	private	Singapore Citizen	English	Mandarin
6	22	F	Chinese	HDB34	Singaporean	Mandarin	English
7	24	M	Chinese	hostel	Malaysian	Mandarin	English
7	23	M	Chinese	HDB34	Singaporean	English	Mandarin

Continued on next page

Table B.1 – continued from previous page

group	age	gender	race	housing	nationality	first language	second language
7	22	M	Chinese	hostel	Chinese	Mandarin	English
7	22	F	Chinese	HDB34	Singaporean	Mandarin	English
7	23	M	Chinese	HDB34	Singaporean	English	Mandarin
7	24	M	Chinese	private	Singaporean	English	Mandarin
7	22	M	Chinese	HDB34	Singaporean	English	Mandarin
7	23	M	Indian	private	Singaporean	English	Mandarin
7	23	F	Chinese	hostel	malaysian	Mandarin	English
7	24	M	Chinese	hostel	Malaysia	Mandarin	English
7	22	F	Chinese	HDB34	singaporean	Mandarin	English
7	25	M	Chinese	hostel	Malaysian	Mandarin	English
7	22	M	Chinese	HDB5	Singaporean	English	Mandarin
7	22	F	Chinese	hostel	Chinese	Mandarin	English
7	23	F	Chinese	HDB5	Singaporean	Mandarin	Mandarin
7	24	M	Chinese	private	Malaysian	English	Mandarin
7	24	M	Chinese	HDB5	Singaporean	English	Mandarin
7	24	F	Chinese	hostel	Malaysia	Mandarin	English
7	23	F	Chinese	HDB34	Singapore Citizen	English	Mandarin
7	25	M	Chinese	private	singaporean	Mandarin	English
8	21	F	Chinese	HDB34	Singaporean	English	Mandarin
8	21	F	Chinese	HDB5	Singaporean	English	Mandarin
8	23	F	Chinese	HDB5	SINGAPOREAN	English	Mandarin
8	21	F	Chinese	private	Singaporean	English	Mandarin
8	24	M	Chinese	HDB34	singaporean	English	Mandarin
8	24	M	Chinese	HDB5	Singaporean	English	Mandarin

Continued on next page

Table B.1 – continued from previous page

group	age	gender	race	housing	nationality	first language	second language
9	24	M	Chinese	HDB5	Singaporean	Mandarin	English
9	24	M	Chinese	HDB34	Singaporean	Mandarin	English
9	23	M	Chinese	hostel	SG	Mandarin	English
9	23	F	Chinese	hostel	Indonesia	Bahasa Indonesia	English
9	23	M	Chinese	HDB5	Malaysian	English	Mandarin
9	24	F	Chinese	HDB34	Singaporean	Mandarin	English
9	23	M	Chinese	HDB5	Singaporean	English	Mandarin

Bibliography

- [Achtert et al., 2008] Achtert, E., Kriegel, H.-P., and Zimek, A. (2008). ELKI: A software system for evaluation of subspace clustering algorithms scientific and statistical database management. In Ludäscher, B. and Mamoulis, N., editors, *Scientific and Statistical Database Management*, volume 5069 of *Lecture Notes in Computer Science*, chapter 41, pages 580–585. Springer Berlin / Heidelberg, Berlin, Heidelberg.
- [Allan et al., 2007] Allan, H., Müllensiefen, D., and Geraint Wiggins (2007). Methodological Considerations In Studies Of Musical Similarity. In *Proceedings of the International Symposium on Music Information Retrieval*.
- [Anglade and Dixon, 2008] Anglade, A. and Dixon, S. (2008). Characterisation of harmony with inductive logic programming. In *Proceedings of the International Symposium on Music Information Retrieval*.
- [Angluin and Laird, 1988] Angluin, D. and Laird, P. (1988). Learning from noisy examples. *Machine Learning*, 2(4):343–370.
- [Arnold et al., 2007] Arnold, A., Nallapati, R., and Cohen, W. W. (2007). A comparative study of methods for transductive transfer learning. In *ICDM Workshop on Mining and Management of Biological Data*.
- [Beguin and Thisse, 1979] Beguin, H. and Thisse, J.-F. (1979). An axiomatic approach to geographical space. *Geographical Analysis*, 11(4):325–341.
- [Benitez et al., 1997] Benitez, J., Castro, J., and Requena, I. (1997). Are artificial neural networks black boxes? *Neural Networks, IEEE Transactions on*, 8(5):1156 –1164.
- [Bertin-Mahieux et al., 2008] Bertin-Mahieux, T., Eck, D., Maillet, F., and Lamere, P. (2008). Autotagger: A model for predicting social tags from

- acoustic features on large music databases. *Journal of New Music Research*, 37:115–135.
- [Breiman and Breiman, 1996] Breiman, L. and Breiman, L. (1996). Bagging predictors. In *Machine Learning*, pages 123–140.
- [Bullock, 2007] Bullock, J. (2007). LibXtract: A Lightweight Library for Audio Feature Extraction. In *Proceedings of the International Computer Music Conference*, Copenhagen, Denmark. ICMA.
- [Burgoyne et al., 2007] Burgoyne, J. A., Pugin, L., Kereliuk, C., and Fujinaga, I. (2007). A Cross-Validated Study Of Modelling Strategies For Automatic Chord Recognition In Audio. In *Proceedings of the International Conference on Music Information Retrieval*.
- [Chapelle et al., 2006] Chapelle, O., Schölkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press.
- [Consortium, 2007] Consortium, T. S. (2007). Smc roadmap.
- [Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.
- [Craft et al., 2007] Craft, A. J. D., Wiggins, G. A., and Crawford, T. (2007). How many beans make five? the consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. In Dixon, S., Bainbridge, D., and Typke, R., editors, *ISMIR*, pages 73–76. Austrian Computer Society.
- [Cross, 2001] Cross, I. (2001). Music, cognition, culture and evolution. *Annals of the New York Academy of Sciences*, 930:28–42.
- [Datta et al., 2006] Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *ECCV (3)*, pages 288–301.
- [DeBruine et al., 2007] DeBruine, L. M., Jones, B. C., Unger, L., Little, A. C., and Feinberg, D. R. (2007). Dissociating averageness and attractiveness: attractive faces are not always average. *J Exp Psychol Hum Percept Perform*, 33(6):1420–1430.

- [Dhanaraj and Logan, 2006] Dhanaraj, R. and Logan, B. (2006). Automatic prediction of hit songs. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 488–491.
- [Dowling and Harwood, 1986] Dowling, W. J. and Harwood, D. L. (1986). *Music Cognition*. Academic Press, San Diego.
- [Duda and Hart, 1973] Duda, R. O. and Hart, P. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.
- [Dudani, 1976] Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-6(4):325–327.
- [Eisenthal et al., 2006] Eisenthal, Y., Dror, G., and Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation*, 18(1):119–142.
- [Freund and Schapire, 1997] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting.
- [Fu et al., 2011] Fu, Z., Lu, G., Ting, K. M., and Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319.
- [Fukunaga and Hayes, 1989] Fukunaga, K. and Hayes, R. R. (1989). Effects of sample size in classifier design. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(8):873–885.
- [Gahegan, 2000] Gahegan, M. (2000). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, 32(2):113–139.
- [Gama, 2004] Gama, J. a. (2004). Functional trees. *Mach. Learn.*, 55(3):219–250.
- [Geleijnse and Korst, 2006] Geleijnse, G. and Korst, J. H. M. (2006). Efficient lyrics extraction from the web. In *ISMIR*, pages 371–372.
- [Gomez et al., 2009] Gomez, E., Haro, M., and Herrera, P. (2009). Music and geography: Content description of musical audio from different parts

- of the world. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 753–758.
- [Gomez and Herrera, 2008] Gomez, E. and Herrera, P. (2008). Comparative analysis of music recordings from western and non-western traditions by automatic tonal feature extraction. *Empirical Musicology Review*, 3(3):140–156.
- [Govaerts and Duval, 2009] Govaerts, S. and Duval, E. (2009). A web-based approach to determine the origin of an artist. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 261–266.
- [G.Peeters, 2008] G.Peeters (2008). A generic training and classification system for mirex08 classification tasks: Audio music mood, audio genre, audio artist and audio tag. In *MIREX*.
- [Grachten et al., 2009] Grachten, M., Schedl, M., Pohle, T., and Widmer, G. (2009). The ismir cloud: A decade of ismir conferences at your fingertips. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 63–68.
- [Grey, 1977] Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustic Society of America*, 61(5):1270–1277.
- [Halberstadt and Rhodes, 2003] Halberstadt, J. and Rhodes, G. (2003). It’s not just average faces that are attractive: computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychon Bull Rev*, 10(1):149–156.
- [Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11.
- [Hamel and Eck, 2010] Hamel, P. and Eck, D. (2010). Learning features from music audio with deep belief networks. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 339–344, Utrecht, The Netherlands. <http://ismir2010.ismir.net/proceedings/ismir2010-58.pdf>.
- [Hanslick, 1891] Hanslick, E. (1891). *The Beautiful in Music*. Novello, Ewer and Co.

- [Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.
- [Hinton, 2009] Hinton, G. E. (2009). Deep belief networks. Scholarpedia.
- [Hinton et al., 2006] Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554.
- [Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic analysis. In Laskey, K. B. and Prade, H., editors, *UAI*, pages 289–296. Morgan Kaufmann.
- [Jean-Julien Aucouturier, 2008] Jean-Julien Aucouturier, E. P. (2008). Introduction from genres to tags: A little epistemology of music information retrieval research. *Journal of New Music Research*, 37(2):87–92.
- [Jeon and Landgrebe, 1992] Jeon, B. and Landgrebe, D. (1992). Classification with spatio-temporal interpixel class dependency contexts. *Geoscience and Remote Sensing, IEEE Transactions on*, 30(4):663–672.
- [Joshi et al., 2011] Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q.-T., Wang, J. Z., Li, J., and Luo, J. (2011). Aesthetics and Emotions in Images. *Signal Processing Magazine, IEEE*, 28(5):94–115.
- [Kalayeh and Landgrebe, 1983] Kalayeh, H. M. and Landgrebe, D. A. (1983). Predicting the required number of training samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:664–667.
- [Kanevski et al., 2008] Kanevski, M., Pozdnoukhov, A., and Timonin, V. (2008). Machine learning algorithms for geospatial data. applications and software tools. In Sánchez-Marrè, M., Béjar, J., Comas, J., Rizzoli, A., and Guariso, G., editors, *iEMSs 2008: International Congress on Environmental Modelling and Software*. International Environmental Modelling and Software Society (iEMSs), Lausanne.
- [Kim et al., 2010] Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 255–266.

- [Knees et al., 2007] Knees, P., Pohle, T., Schedl, M., and Widmer, G. (2007). A music search engine built upon audio-based and web-based similarity measures. In *In Proceedings of Special Interest Group on Information Retrieval (SIGIR)*, pages 447–454, New York, NY, USA.
- [Knees et al., 2005] Knees, P., Schedl, M., and Widmer, G. (2005). Multiple lyrics alignment: Automatic retrieval of song lyrics. In *ISMIR*, pages 564–569.
- [Kohonen, 1982] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.
- [Konecni et al., 2008] Konecni, V. J., Brown, A., and Wanic, R. A. (2008). Comparative effects of music and recalled life-events on emotional state – konecni et al. 36 (3): 289 – psychology of music. *Psychology of Music*, 36(3):289–308.
- [Kotsiantis et al., 2006] Kotsiantis, S., Zaharakis, I., and Pintelas, P. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190.
- [Kruskal, 1964] Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.
- [Landwehr et al., 2005] Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. *Machine Learning*, 59(1-2):161–205.
- [Last.fm, 2013a] Last.fm (2013a). Last.fm charts: Top tags. [Online; accessed 27-April-2013].
- [Last.fm, 2013b] Last.fm (2013b). Last.fm social media music site. [Online; accessed 27-April-2013].
- [Laurier et al., 2008] Laurier, C., Grivolla, J., and Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *Proceedings of the Seventh International Conference on Machine Learning and Applications*, pages 688–693.
- [Levitin, 2006] Levitin, D. (2006). *This is Your Brain on Music*. Atlantic Books.
- [Li et al., 2007] Li, Q., Myaeng, S. H., and Kim, B. M. (2007). A probabilistic music recommender considering user opinions and audio features.

- Information Processing & Management*, 43(2):473 – 487. jce:titlejSpecial issue on AIRS2005: Information Retrieval Research in Asiaj/ce:titlej.
- [Li and Ogihara, 2003] Li, T. and Ogihara, M. (2003). Detecting emotion in music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) 2003*.
- [Lichte, 1941] Lichte, W. H. (1941). Attributes of complex tones. *Journal of Experimental Psychology*, 28:455–480.
- [Liu et al., 2009] Liu, Y., Xiang, Q., Wang, Y., and Cai, L. (2009). Cultural style based music classification of audio signals. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*.
- [Magno and Sable, 2008] Magno, T. and Sable, C. (2008). A Comparison Of Signal-Based Music Recommendation To Genre Labels, Collaborative Filtering, Musicological Analysis, Human Recommendation, And Random Baseline. In *Proceedings of the International Symposium on Music Information Retrieval*.
- [Manaris et al., 2003] Manaris, B. Z., Vaughan, D., Wagner, C., Romero, J., and Davis, R. B. (2003). Evolutionary music and the zipf-mandelbrot law: Developing fitness functions for pleasant music. In *EvoWorkshops*, pages 522–534.
- [Mandel and Ellis, 2008] Mandel, M. and Ellis, D. P. (2008). A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165.
- [Matthews, 1896] Matthews, G. B. ((1896)). On the partition of numbers. *Proc. London Math. Soc.*, s1-28(1):486–490.
- [McDermott and Hauser, 2005] McDermott, J. and Hauser, M. (2005). The origins of music: Innateness, uniqueness, and evolution. *Music Perception*, 23(1):29–59.
- [McEnnis et al., 2005] McEnnis, D., McKay, C., and Fujinaga, I. (2005). Jaudio: A feature extraction library. In *Proceedings of the International Conference on Music Information Retrieval*.
- [McKay and Bainbridge, 2011] McKay, C. and Bainbridge, D. (2011). A musical web mining and audio feature extraction extension to the greenstone

- digital library software. In Klapuri, A. and Leider, C., editors, *ISMIR*, pages 459–464. University of Miami.
- [McKay and Fujinaga, 2004] McKay, C. and Fujinaga, I. (2004). Automatic genre classification using large high-level musical feature sets. In *International Conference on Music Information Retrieval*, pages 525–530.
- [McKay and Fujinaga, 2006] McKay, C. and Fujinaga, I. (2006). Musical genre classification: Is it worth pursuing and how can it be improved? In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- [Melville and Mooney, 2004] Melville, P. and Mooney, R. J. (2004). Creating diversity in ensembles using artificial data. *Journal of Information Fusion: Special Issue on Diversity in Multi Classifier Systems*, 6(1):99–111.
- [Mierswa et al., 2006] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. In Ungar, L., Craven, M., Gunopulos, D., and Eliassi-Rad, T., editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA. ACM.
- [Miller, 1966] Miller, R. (1966). *Simultaneous statistical inference*. McGraw-Hill series in probability and statistics. McGraw-Hill.
- [Mitchell, 1980] Mitchell, T. M. (1980). The need for biases in learning generalizations. Technical report, Rutgers University, New Brunswick, NJ.
- [Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.
- [Muggleton and de Raedt, 1994] Muggleton, S. and de Raedt, L. (1994). Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19:629–679.
- [Navada et al., 2011] Navada, A., Ansari, A., Patil, S., and Sonkamble, B. (2011). Overview of use of decision tree algorithms in machine learning. In *Control and System Graduate Research Colloquium (ICSGRC), 2011 IEEE*, pages 37–42.

- [Nettleton et al., 2010] Nettleton, D., Orriols-Puig, A., and Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33:275–306.
- [Pachet and Aucouturier, 2004] Pachet, F. and Aucouturier, J.-J. (2004). Improving timbre similarity: How high is the sky? *Journal of negative results in speech and audio sciences*, 1(1):1–13.
- [Pachet and Roy, 2008] Pachet, F. and Roy, P. (2008). Hit song science is not yet a science. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- [Paiement et al., 2008] Paiement, J.-F., Grandvalet, Y., and Eck, D. (2008). A distance model for rhythms. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki.
- [Park et al., 2012] Park, D. H., Kim, H. K., Choi, I. Y., and Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications*, 39(11):10059 – 10072.
- [Pearl, 1985] Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine*, pages 329–334.
- [Pearson, 1901] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine (S6)*, 2(11):559–572.
- [Photo.net, 2013] Photo.net (2013). Photo.net online photo sharing. [Online; accessed 27-April-2013].
- [Proutskova, 2007] Proutskova, P. (2007). Musical memory of the world data infrastructure in ethnomusicological archives. In *ISMIR*.
- [Q and King, 2013] Q, C. and King, R. D. (2013). Machine learning as an objective approach to understanding music. In Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., and Ras, Z., editors, *New Frontiers in Mining Complex Patterns*, volume 7765 of *Lecture Notes in Computer Science*, pages 64–78. Springer Berlin Heidelberg.
- [Q, 2008] Q, C. E. (2008). Tacet: An automatic modular system for music genre classification. Master’s thesis, Aberystwyth University.

- [Rebello et al., 2010] Rebello, A., Capela, G., and Cardoso, J. S. (2010). Optical recognition of music symbols - a comparative study. *IJDAR*, 13(1):19–31.
- [Rhodes et al., 2001] Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., and Akamatsu, S. (2001). Attractiveness of facial averageness and symmetry in non-western cultures: in search of biologically based standards of beauty. *Perception*, 30(5):611–625.
- [Ripley, 2005] Ripley, B. (2005). *Spatial Statistics*. Wiley Series in Probability and Statistics. Wiley.
- [Rodriguez et al., 2006] Rodriguez, J., Kuncheva, L., and Alonso, C. (2006). Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619–1630.
- [Sammon, 1969] Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 18(5):401–409.
- [Schedl et al., 2007] Schedl, M., Widmer, G., Pohle, T., and Seyerlehner, K. (2007). Web-based detection of music band members and line-up. In *ISMIR*, pages 117–118.
- [Schellenberg et al., 2008] Schellenberg, G., Peretz, I., and Viellard, S. (2008). Liking for happy and sad sounding music: Effects of exposure. *Cognition and Emotion*, 22:218–237.
- [Schneier, 2011] Schneier, B. (2011). Schneier’s law. [Online; accessed 17-September-2012].
- [Sébastien et al., 2012] Sébastien, V., Ralambondrainy, H., Sebastien, O., and Conruyt, N. (2012). Score analyzer: Automatically determining scores difficulty level for instrumental e-learning. In Gouyon, F., Herrera, P., Martins, L. G., and Müller, M., editors, *ISMIR*, pages 571–576. FEUP Edições.
- [SEDAC and CIAT, 2012] SEDAC, C. U. and CIAT (cited August 2012). Ciesin, columbia university; and ciat: Gridded population of the world, version 3 (gpwv3). <http://sedac.ciesin.columbia.edu/gpw>.
- [Sessions, 1970] Sessions, R. (1970). *Questions about music*. Harvard University Press, Cambridge, Mass.

- [Settles, 2009] Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.
- [Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.
- [Siegel and Castellan, 1988] Siegel, S. and Castellan, N. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw–Hill, Inc., second edition.
- [Sigurdsson et al., 2006] Sigurdsson, S., Petersen, K. B., and Lehn-Schißler, T. (2006). Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music. In *Proceedings of the 7th International Symposium on Music Information Retrieval (ISMIR)*.
- [Skowronek et al., 2007] Skowronek, J., McKinney, M., and van de Par, S. (2007). A Demonstrator For Automatic Music Mood Estimation. In *Proceedings of the International Symposium on Music Information Retrieval*.
- [Sugimoto et al., 2010] Sugimoto, T., Kobayashi, H., Nobuyoshi, N., Kiriya, Y., Takeshita, H., Nakamura, T., and Hashiya, K. (2010). Preference for consonant music over dissonant music by an infant chimpanzee. *Primates*, 51(1):7–12.
- [Szpunar et al., 2004] Szpunar, K., Schellenberg, E., and Pliner, P. (2004). Liking and memory for musical stimuli as a function of exposure. *J Exp Psychol Learn Mem Cogn*, 30(2):370–81.
- [Thompson, 2007] Thompson, S. (2007). Determinants of listeners’ enjoyment of a performance. *Psychology of Music*, 35(1):20–36.
- [Turnbull et al., 2009a] Turnbull, D., Barrington, L., Lanckriet, G. R. G., and Yazdani, M. (2009a). Combining audio content and social context for semantic music discovery. In Allan, J., Aslam, J. A., Sanderson, M., Zhai, C., and Zobel, J., editors, *SIGIR*, pages 387–394. ACM.
- [Turnbull et al., 2009b] Turnbull, D. R., Barrington, L., Lanckriet, G., and Yazdani, M. (2009b). Combining audio content and social context for semantic music discovery. In *Proceedings of ACM Special Interest Group on Information Retrieval (SIGIR)*, pages 387–394, New York, NY, USA.

- [Tzanetakis and Cook, 2000] Tzanetakis, G. and Cook, P. (2000). Marsyas: a framework for audio analysis. *Organised Sound*, 4(3):169–175.
- [Tzanetakis and Cook, 2002] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.
- [Tzanetakis et al., 2007] Tzanetakis, G., Kapur, A., Schloss, A., and Wright, M. (2007). Computational ethnomusicology. *Journal of Interdisciplinary Music Studies*, 1(2):1–24.
- [Vapnik, 1998] Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.
- [Widmer et al., 2005] Widmer, G., Dixon, S., Knees, P., Pampalk, E., and Pohle, T. (2005). From sound to sense via feature extraction and machine learning: Deriving high-level descriptors for characterising music.
- [Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [Witvliet and Vrana, 2007] Witvliet, C. and Vrana, S. (2007). Play it again sam: Repeated exposure to emotionally evocative music polarizes liking and smiling responses, and influences other affect reports, facial emg, and heart rate. *Cognition and Emotion*, 21:3–25.
- [Wu and Takatsuka, 2005] Wu, Y. and Takatsuka, M. (2005). The geodesic self-organizing map and its error analysis. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science*, volume 38, pages 343–351.